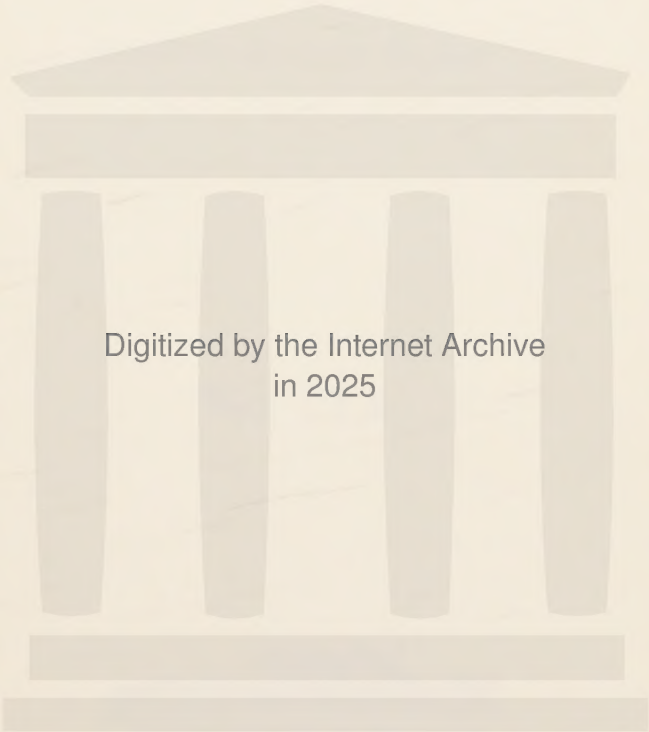

Columbia University

Contributions to Education

Teachers College Series

No. 273

AMS PRESS
NEW YORK

THE IMPROVEMENT OF INTELLIGENCE TESTING

BY

HAROLD H. ABELSON, PH.D.

TEACHERS COLLEGE, COLUMBIA UNIVERSITY
CONTRIBUTIONS TO EDUCATION, NO. 273

BUREAU OF PUBLICATIONS
Teachers College, Columbia University
NEW YORK CITY
1927

Library of Congress Cataloging in Publication Data

Abelson, Harold Herbert, 1904-
The improvement of intelligence testing.

Reprint of the 1927 ed., issued in series: Teachers College, Columbia University. Contributions to education, no. 273.
Originally presented as the author's thesis, Columbia.
Bibliography: p.
1. Mental tests. I. Title. II. Series: Columbia University. Teachers College. Contributions to education, no. 273.
BF431.A422 1972 153.9'3 76-176502
ISBN 0-404-55273-0

Reprinted by Special Arrangement with Teachers College Press, New York, New York

From the edition of 1927, New York
First AMS edition published in 1972
Manufactured in the United States

AMS PRESS, INC.
NEW YORK, N. Y. 10003

FOREWORD

Education may in a sense be regarded as the harnessing for individual and social purposes of the stupendous potentiality lodged in the brain of each person. The highest realization of the intellectual capacity of each individual is dependent to a large extent upon the rapidity and accuracy with which a knowledge of that capacity may be ascertained. The essential objective of research in mental measurements has been to increase the valid discriminatory power of tests without incurring an uneconomical increase of testing time. The present study offers a relatively new approach to the problem of test improvement. It represents a tendency toward the selection and evaluation of test material in more and more minute and elementary units, on the one hand, and in terms of relatively more objective criteria, on the other. It signifies a movement toward the scientific method that may lead, it is hoped, to outstanding and perhaps startling improvements in psychological testing.

Those who have followed certain of the recent developments in the work of Professor William A. McCall will clearly discern how this study merely carries into a new field many of his thoughts on item analysis. Only those who have worked with McCall and have consequently felt the influence of his friendship and stimulating advice, can know the full significance of my debt to him. The professors and students of the psychological seminar of Teachers College, Columbia University, have made many important suggestions. Professors Rudolf Pintner and Ben Wood have given valued criticisms of the initial plans of the study. To Professor Henry A. Ruger I am indebted for his very generous and helpful statistical advice. The criticism of a section of the work by Professor Edward L. Thorndike proved very enlightening. Conference with Miss Harriet Barthelmess, coworker in the field, resulted in the clarification of certain important issues.

A group of persons associated with the College of the City of New York are responsible for much help in the facilitation

of the investigation. To Professor Paul Klapper, Dean of the School of Education of the College, I am indebted for many instances of hearty and friendly coöperation. Professors Samuel B. Heckman, J. Carleton Bell, and Egbert M. Turner of the Department of Education of the College, rendered valuable service in connection with the selection and the administration of the tests employed. The willing assistance of Professor Morton Gottschall, Registrar of the College, facilitated the collection of data on the college achievement of the students tested. The progress of the work was materially accelerated by the use of the initial test results determined by the writer in connection with a study conducted under the joint auspices of the School of Education and the School of Business and Civic Administration, of the College, for which help I wish to acknowledge my appreciation.

It is very difficult to express adequately my gratefulness to my wife, Lucie Bernard Abelson, for her ample assistance and encouragement throughout the course of the study.

H. H. A.

CONTENTS

CHAPTER		PAGE
I	The Problem	1
	Problem and Approach	1
	Hypothetical Discussion of Problems	1
II	The Data and Their Initial Treatment	4
	Selection of Items for Study	4
	Administration of Items to College Entrants	4
	Determination of College Success Criterion Scores	7
	Tabulations and Initial Computations	10
III	The Improvement of Scoring Through Item Analysis	14
	Technique for Assigning the New Values to Item Responses	14
	Results and Their Interpretation	15
	Empirical Comparison of the Old with the New Scoring Methods	28
	Tentative Trial of Modifications of the Method of Determining New Scoring Values	33
IV	The Problem of the Choice of the Best Items	37
	Choice of the Item Coefficient	37
	Characteristics of the Item Coefficient	39
	Empirical Study of the Reliability of the Item Coefficient	43
	Practical Effectiveness of the Item Coefficient in Choosing the Best Items	46
	Modification of the Item Coefficient	50
	Determination of the Objective Factors Associated with Item Goodness	51
V	The Analysis of the Subtests	56
	Determination of the Measures Employed	56
	Results	59
VI	Summary and Conclusions	61
APPENDIX		
I	Suggestions for Decreasing the Labor Associated with Item Analysis	65
II	Miscellaneous Supplementary Results	67
III	Bibliography	69

THE IMPROVEMENT OF INTELLIGENCE TESTING

CHAPTER I

THE PROBLEM

PROBLEM AND APPROACH

How may intelligence testing be improved? Of the several possible ways the following are selected for study in the present investigation:

1. The better scoring of responses to item stimuli.
2. The better choice of items.

In the past, barring a few exceptional studies, items have been selected and responses evaluated on a highly subjective basis. Here, more objective techniques are attempted. Moreover, while in the development of mental measurements, entire examinations and subtests have been studied statistically with profit, any outstanding improvement of our present instruments must come as a result of the analysis of the item and of responses within the item. The keynote of the present study is, then, objective item analysis. College entrance testing is made focal, but suggestive applications to all levels and to all kinds of testing may be made.

HYPOTHETICAL DISCUSSION OF PROBLEMS

The writer has found it both stimulating and clarifying to ask his research students, when they had selected their problems, the question, "With infinite but human resources how would you solve your problems?" The application of this question to the problems on hand many prove helpful, first, in presenting a broad outline of the work of the investigation; and second, as making possible a clearer enunciation of the assumptions and limitations involved.

With unlimited facilities, then, how might one determine the best scoring of item responses to obtain the optimum college success prediction? One answer follows:

1. Tentatively construct or select an infinite number and variety of test items.

2. Administer these items to a large sampling of college entrants, equating for each item such factors as fatigue, placement of the item, mental set, and so on.

3. Determine the success in college of each student after four years (or possibly his later success in using college training toward life adjustment).

4. Analyze the possible responses to the stimulus of each item. (These responses might be classified under several types.)

5. Compute the average degree of college success achieved by the students, grouped, for each item separately, according to their response (or type of response).

6. Assign to each response (or type of response) the average college success score computed for it.

7. Assuming that items are selected with due regard to their intercorrelations, this technique would theoretically give the optimum predictive scoring to be used with subsequent similar groups of students.

And similarly the hypothetical solution of the problem of the best choice of items might be as follows:

1, 2, and 3 as above.

4. Assuming the best scoring of items as indicated above, determine the correlation between responses to each item and the criterion of college success, summarizing this relationship in a coefficient for each item.

5. Determine the intercorrelations of each item with every other item.

6. By the application of multiple regression formulæ, it would then be possible to select the best group of items for the desired prediction of college success, provided, of course, that subsequent entrant groups were similar to the experimental group.

The above analyses present in a rough way, free from specific considerations and also from practical limitations, possible answers to the problems studied. Any practical and specific attempt at solution must carry with it several modifications of the theoretically sound solutions, thereby increasing the number of

assumptions involved and lessening the probable effectiveness of a new method of selecting and scoring test items. The following chapters explain essentially the various methods attempted and the results obtained as regards the improvement of the tests studied.

CHAPTER II

THE DATA AND THEIR INITIAL TREATMENT

SELECTION OF ITEMS FOR STUDY

With the coöperation of members of the Department of Education of the College of the City of New York, five leading college entrance intelligence examinations were selected to be administered during the fall of 1925 to the incoming freshman class. The chosen examinations were:

The Thorndike Intelligence Examination for High School Graduates.

The Roback Mentality Tests for Superior Adults.

The Brown University Psychological Examination.

The American Council on Education Psychological Examination, 1924 Edition.

Thurstone's Psychological Examination IV (1919).

After these had been administered as described in this chapter and their subtests had been analyzed as explained in Chapter V, the items of certain subtests were selected for intensive analysis. The selection was made with the purpose of obtaining a variety of test types. Tabulated descriptions of the tests studied are presented on page 16, Chapter III. In all, 205 of the approximate 450 items contained in the above examinations were subjected to intensive item analysis.

ADMINISTRATION OF ITEMS TO COLLEGE ENTRANTS

The 621 lower freshmen who had entered the day session of the College of the City of New York during September, 1925, were employed as subjects. They were divided into four groups, hereafter termed A, B, C, and D. Groups A, B, and C were selected alphabetically. The D group consisted of those who were absent from the regular testing period. This group was tested some five weeks later. Table I gives a tabulated description of the groups and of the examinations employed with each.

Groups A, B, and C are held to be about as comparable as samples selected by chance. Group D includes, perhaps, a small number of students who may have attempted to avoid the tests. This may account for the lower average C.S.C. Score and the greater variability of this group. Since greater variability tends to raise coefficients of correlations, a slight downward correction would have to be made in comparing a coefficient based on this group with those based on the other groups.

TABLE I

DATA ON THE EXPERIMENTAL GROUPS

Group	No. of Students	Examinations Taken	Time in Minutes	Mean of C.S.C. Scores [1]	S. D. of C.S.C. Scores
A	175	Thorndike	170	50.57	9.99
B	248	1. Brown	70		
		2. McCall Multi-mental (Experimental Form)	45	49.88	9.90
			115		
C	137	1. American Council Test 1 (Completion)	10		
		2. Roback (Test 2 and 8 omitted) ..	130	49.53	9.30
			140		
D	61	1. Thurstone IV	30		
		2. American Council Tests	50	48.38	11.64
			80		

[1] See pp. 7 ff. for a full account of the derivation of these scores.

Application to outside groups of the results found with the subjects employed may be limited by certain special characteristics of the experimental group.

The group is highly selected along linguistic and academic lines. Admission to the day session of the College is limited (barring special examinations rarely passed) to those who have maintained an average high school mark of at least 75 per cent. These marks are teachers' marks and not Regents examination grades. They are, according to authentic estimation, from five

to ten percentage points lower than the Regents' grades. That the selection is strenuously exclusive is indicated by the fact that roughly 40 or 50 per cent of the applicants of the two preceding semesters had been refused admission to the day session of the college.

Quantitative evidence as regards the status of the experimental group and its variability in mental functions is made possible by means of comparisons with other colleges which participated in the testing program of the American Council on Education for the academic year of 1924-25. During this year about 950 freshmen at the College of the City of New York took the American Council tests. The ranks of this group, compared with 59 other colleges, first rank being assigned to the highest average score, and so on, ranged between first and eleventh. A comparison between the 1924 group and the 1925 group employed in this study indicates a still higher selection in the case of the latter. The mean score on the American Council Completion of the 1925 group of freshmen was 22.18, which is slightly more than the mean of the scores of the 1924 group. The explanation for the higher score of the more recent group is very likely the fact that the entrance requirements were raised and were more stringently enforced.

Since athletics and social life are minimized in this college as compared with other colleges, a more studious type of individual tends to seek entrance. Moreover, an unusually large proportion must engage in outside work in order to earn their way through college. Although relatively few are foreign-born, a large proportion are of foreign-born parentage. Practically all have their homes in New York City, and the larger percentage of them have lived in an urban community all their lives.

While, with the exception of the D group, the groups were tested simultaneously and while the usual uniform test conditions were maintained, it was naturally impossible to eliminate such factors as item placement, time per item, mental set, fatigue, practice effect, and the like. Hence certain items might have been given the advantage over others as predictive agents because of the favorable operation of some one or some combination of these factors. Except for the omission of several subtests, as indicated, the directions of the respective test authors were followed literally.

DETERMINATION OF COLLEGE SUCCESS CRITERION SCORES

The determination of measures of success at college necessarily falls far short of what one might do with unlimited resources. The truest criterion according to the best current educational theory would be the extent to which college education had modified the individual so that he might better adjust himself to the complete environment. However, the only fairly objective measurements available are the subject grades which the students earn. In spite of the wide disparity between the fundamental criterion mentioned above and the college grades, the latter have been employed, and with the following partial, if not complete, justification:

1. They are the only feasible measures.
2. They are used at present as the basis of promotion, acceleration, graduation, honoring of students, etc.
3. They are in line with the present aim of the college, academic and restricted, perhaps, as it may be.
4. The present study can aim only at the clarification and the development of techniques, rather than at the determination of ultimate results.
5. The criterion employed undoubtedly correlates fairly highly with the truest criterion, and since the comparisons of items rather than the determination of absolute values are aimed at, it is quite likely that the results would remain virtually the same with respect to the truest criterion.

The grades of the first semester alone were employed because:

1. Previous reports [1] of correlations between tests and college grades have shown no increase, or only a moderate increase, in correlations when grades of additional semesters have been added in the computation of the criterion score.
2. A large proportion leave college after the first semester, thereby further restricting the group or necessitating many subjective judgments as regards college success or failure.
3. The utilization of grades of additional semesters would have entailed a delay in the work which neither was feasible nor seemed justifiable as indicated above.

In addition to the grades of each student in each subject which he had attempted, the amount of work attempted in terms of

[1] The titles of several such reports are listed in the Bibliography.

credits was also included in the computation of the criterion. Results of a questionnaire study investigating into the number of hours of outside work, the amount of study for college work, and the like, were not used because of the high subjectivity and the apparent inaccuracy of the responses. The problem resolved itself, therefore, into the following aspects:

1. How to equate marks given in different courses, having standards varying in severity.

2. How to combine the average equated grade with the number of credits attempted, or, in other words, how to combine the quality with the quantity of college achievement.

The marks were equated by means of the so-called T scale technique, that is, by reducing the distribution of the marks for each college subject to a single form, that of the normal frequency curve. T values for the various letter grades are given in Table II.

TABLE II

THE T VALUES FOR THE VARIOUS COLLEGE GRADES EARNED IN VARIOUS COURSES

	A	B	C	D	E	F	(F)	N
Art 1	64	55.5	48.5	43		37	30.5	172
Chemistry 1	67	56	47.5	38		30.5	25	150
Chemistry 1*a*	70	61	52.5	43.5		34.5	26	108
Economics 1, 2	73.5	61.5	50.5	41	35	30		108
English 1	72	58.5	47	34.5		27.5	24	339
English 3	67	58.5	49.5	40.5	34	30.5		181
French 1, 2, 3, 4, 53, 54	68.5	60	52.5	44	36.5	33	25	249
French 51	62	53	44	35		28		33
German 1, 3	66	57	45	34			27.5	38
Government and History	71	59.5	50	42	41.5	30.5		60
Hygiene 1	79.5	62.5	52	41.5	36.5	35	20.5	581
Latin 1, 3	68	58	52	46	40	37	33	70
Latin 51, 53	68.5	58.5	50	42.5	36.5	33	25	90
Mathematics 1, 2, 3, 4, 7, 53, 1-2	65.5	58.5	52.5	46.5	41.5	39.5	36.5	461
Military Science 1	61	54	44.5	39	34	30		551
Physics 1, 2	65	59.5	53	43.5	35	32.5	28.5	99
Philosophy 1		67	55	42.5				11
Public Speaking 1, 2, 1-2	69	59	50	41	34	30.5	22	358
Spanish 1, 3	67.5	60	52	42.5		34	29	84

The symbol (F) signifies the forced or voluntary dropping of a course by a student. E signifies a condition; F a failure.

To facilitate computations, these values were reduced to a scale of from 0 through 11, according to the following equivalents:

T Score: from	20	25	30	35	40	45	50	55	60	65	70	75
to	24.5	29.5	34.5	39.5	44.5	49.5	54.5	59.5	64.5	69.5	74.5	79.5
Reduced Scale Value	0	1	2	3	4	5	6	7	8	9	10	11

The numerical, equated, and reduced grades for each student were averaged, grades received in less important courses being given half weight. These averages gave the "quality" ratings. Each course has a given credit value, one credit signifying the expectancy of two and one-half hours of work weekly on the part of the student. Simply summing the number of credits attempted by each student resulted in the "quantity" ratings. Since students varied in the amount of work undertaken, the quality ratings alone did not seem adequate in offering a fair judgment of the students' work. The two ratings were weighted so as to give the best prediction score on the Thorndike Examination Part I. In order to obtain these weights the following measures were computed (with a sampling of 100 cases):

Coefficients of Correlation		Standard Deviations	
Thorndike Part I with Quality Ratings	(r 12) .226	(σ1)	Thorndike Part I 27.1
Thorndike Part I with Quantity Ratings	(r 13) .121	(σ2)	Quality Ratings 10.5
Quality Ratings with Quantity Ratings	(r 23) .110	(σ3)	Quantity Ratings 1.3

These values were substituted in the formulæ:

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{(1 - r_{13}^2)^{\frac{1}{2}}\,(1 - r_{23}^2)^{\frac{1}{2}}}\,;\quad r_{13.2} = \frac{r_{13} - r_{12}\, r_{23}}{(1 - r_{12}^2)^{\frac{1}{2}}\,(1 - r_{23}^2)^{\frac{1}{2}}}$$

$$\text{and } \sigma^2_{1.23} = \sigma_1^2\,(1 - r^2_{12})\,(1 - r^2_{13.2})\,;\ \text{etc.}$$

$$\text{and } b_{12.3} = r_{12.3}\,\frac{\sigma_{1.23}}{\sigma_{2.13}}\,;\ \text{etc.}$$

$$\text{to give: } x_1 = b_{12.3}\, x_2 + b_{13.2}\, x_3,$$

the regression formula for predicting x_1 from a knowledge of x_2 and x_3, $b_{12.3}$ and $b_{13.2}$ being the respective weights of the last named variables. The x_1, x_2 and x_3 refer to deviations from the respective means of the Thorndike scores, the quality ratings and the quantity ratings.

The values of $b_{12.3}$ and $b_{13.2}$ are .57 and 2.3 respectively. The weights actually assigned were about 2 to the quantity and 7 to

the quality ratings. The resulting composite scores were T scaled, employing the four groups. These T scores were converted into plus and minus deviations from the general mean. The appropriate deviation was then entered upon a narrow strip next to the number representing each student.

Twenty-three students had left college so that the usual ratings were not available for them. Nine had been dropped; ten had left late in the term; and four had left early or with no indication on their records as to the time of leaving. Several persons were asked to place these three types on a percentile scale in comparison with the remaining students. The percentile rankings averaged for the respective groups approximately 5th, 20th, and 30th. The T values corresponding to these percentiles are respectively, 36, 43, and 47. The deviations from the mean of 50 are consequently — 14, — 7 and — 3. These special scores were treated as similar to the other C.S.C. Scores.

Results found with the use of the C.S.C. Scores are subject to the partial inconsistency of these scores. An approximate measure of their reliability was obtained by dividing the grades for each student into two parts, computing the quality rating for each part and determining the coefficient of correlation between the ratings of the two divisions. The resultant coefficient with a sampling of 100 students proved to be .388. But the coefficient thus found represented the consistency between the halves of the C.S.C. Scores. By means of the Spearman-Brown formula, the probable reliability coefficient of one whole scole (i.e., the score for a whole semester) with a second whole score was computed. It proved to be .559. A test perfectly designed to predict the true C.S.C. Score, i.e., the average score of an infinite number of single scores, would, within chance variations, yield a correlation coefficient of only .559 with the single scores used in this study. This is obvious, since an external test can hardly predict test scores of a trait better than do similar scores on the same trait. The low validity coefficients quoted below are more easily explained in the light of the above fact.

TABULATIONS AND INITIAL COMPUTATIONS

Response Analysis.—The fourth step in the hypothetical solution of page 2 calls for the analysis of the significantly different possible responses to each item stimulus. Several considera-

tions operated in this analysis. First, it was necessary to obtain the scores of the student according to the test author's method of scoring. Hence, responses credited by the author were retained under his classification. For ease in summing item scores to obtain test scores, numerical symbols corresponding to the number of points allowed in the scoring manual were assigned to these responses. Second, it was advisable to keep the number of response types low. Therefore, a response likely to occur rarely, as determined by a cursory inspection, or one only slightly distinctive, was classed with the type of response most similar to it. Third, in connection with tests too difficult to be finished by a considerable number in the time allotted, it was thought advisable to distinguish between two kinds of omissions; i.e., omissions in the middle of the test followed by an attempt later on in the same test, as distinguished from omissions not followed by a later attempt. The difference involved is that in the first case the student, after attempting the item, was unable to make a satisfactory response, while in the second case, the student, whether able or not to answer correctly, had insufficient time apparently to attempt the item. Fourth, it was held to be worth investigation to determine whether responses scored "wrong" by the author were of varying value. Hence, where possible and feasible, attempts rated "wrong" were classified under several subgroups. Fifth, especially in the case of subjectively scored tests it was thought advisable to provide for doubtful responses. In general, the fourth and fifth considerations were subordinated to the second.

Tabulation of Response Symbols for Each Item and for Each Student.—The effective study of item responses is largely dependent upon the efficiency of the tabulation methods employed. The tabulation techniques here used are presented with some detail, though briefly, in order that the principles underlying the new scoring device might be more concretely understood.

The tabulation form is illustrated in Figure 1. The numbers assigned to the students are listed vertically at the extreme left. The item numbers for the subtest studied are listed horizontally at the top. Thus in the figure 20 students and 40 items are represented. The symbols, 1, *a*, *X*, etc., indicate the response of a student to the various item situations. The scorer, as he reads the test booklet, writes the appropriate symbol in the

| Student | Item | | |

appropriate square. This method saves a considerable amount of time, making possible the scoring and the tabulation of responses in only slightly more time than the scoring according to the usual method ordinarily takes. Referring to the illustration, any one student's responses to the various items may be noted by glancing along the row allotted to him, while the responses made by the students to any one item may be studied by regarding the column assigned to it.

The completed tabulation makes readily available the numerous diversified calculations employed in the study.

CHAPTER III

THE IMPROVEMENT OF SCORING THROUGH ITEM ANALYSIS

TECHNIQUE FOR ASSIGNING THE NEW VALUES TO ITEM RESPONSES

By the "new" scoring method employed in this chapter is meant the technique for assigning objective item response values as described below. The word "new" is applied for want of a convenient name. This chapter treats especially of the technique which at the outset seemed most promising. Several other methods are briefly discussed at the close of the chapter.

The hypothetical discussion of the problem offered on page 2, Chapter I, roughly outlined the steps to be followed in the determination of the best scoring technique. In Chapter II the procedure was carried through step 4 of the initial outline. There it was indicated how the experimental items were selected and to whom and how they were administered, how the possible responses to each item were classified, how symbols representing the various types of response were tabulated for each student and each item and how, finally, the College Succession Criterion Scores were computed and tabulated for each student.

Considering a single item at a time, the final steps in the process are simply: (1) to compute the mean of the C.S.C. Scores of students grouped for each item according to the various response types of that item. Thus the C.S.C. Scores of all those who made response A of Item 1 are averaged, and so on; and (2) to assign to each response a value equivalent to the mean of the C.S.C. Scores associated with it.

The method is further explained in the Appendix for those who plan to conduct similar work. A simplified illustration is presented here to show the general underlying principle.

Suppose Test A to be a true-false information test. Three kinds of responses are possible: a correct response (according to the author's scoring), an incorrect response, and an omission.

These responses are termed in order: R, W, and X. Suppose ten students took Test A. Assume further that their C.S.C. Scores (not in terms of T deviations in this example) were from lowest to highest, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5. For Item 1 of Test A the results might have appeared as follows:

Student	Response	C.S.C. Score
1	R	4
2	W	2
3	X	3
4	R	5
5	X	2
6	W	1
7	R	3
8	X	4
9	W	2
10	W	3

Summary: Response R of Item 1 was made by three students having C.S.C. Scores of 4, 5, and 3. The mean of the C.S.C. Scores equals 4.00.

Response W was made by four students having C.S.C. Scores of 2, 1, 2, and 3. The mean of the C.S.C. Scores equals 2.00.

Response X was made by three students having C.S.C. Scores of 3, 2, and 4. The mean of the C.S.C. Scores equals 3.00.

Therefore, considering Item 1 by itself, to predict the C.S.C. Score, theoretically, the best values to assign to Responses R, W, and X are 4, 2, and 3 respectively.

The scoring values by this method are determined, then, in terms of the available measure of the function it is desired to predict. This is one of the significant principles which McCall employed in the construction of his Multi-mental Scale. In that test no one of the various possible responses can be said to be correct absolutely; each is right or wrong to a degree. The tests employed in this study involve responses, which, according to the author's scoring, are for the most part planned to be entirely right or entirely wrong; in a few cases only are degrees of worth assigned.

RESULTS AND THEIR INTERPRETATION

How do the scoring values determined by the new technique actually compare with the values assigned by the test authors?

Results have been computed with subtests described in Table III.

The American Council Completion items are of the type:

A (An) ..... (8) is an unmarried .....(5).

The numbers in parentheses indicate the number of letters contained in the respective omissions. The test calls for speedy responses and offers a variety in difficulty and in content. The new scoring response values are tabulated in Table IV. Each blank to be filled in is regarded as an item. The forty items are numbered vertically at the left. There are four types of

TABLE III

SUBTESTS EMPLOYED IN THE COMPARISON OF THE OLD WITH THE NEW SCORING VALUES

Name	Form	Content	Time in Minutes	Number of Items	Number of Cases	Group Employed	Validity Coefficient
American Council 1	Completion	General	10	40	137	B	.252
Thorndike IIG8	True-false	Academic Information	13	60	100	C	.099
Thorndike IIG3	Picture Completion	Mechanical	3	10	100	C	.045
Brown 1	Completion	General	13	20	100	A	.230
Brown 3	Multiple Choice of 8	Opposites	8	20	100	A	.162
Roback 1	Write Word	Abstraction	10	30	68	B	.020
Roback 7	Write Word	Opposites	10	25	68	B	—.052

responses: type 1, the correct word according to the author's scoring key; type *a*, the insertion of a wrong word, of part of a word or of more than one word; type *X*, an omission followed by an attempt later on in the same test; and type *y*, an omission with no later attempt. The *d*1 column gives the amount and direction of the difference between the mean of the C.S.C. Scores of the entire Group B and the average C.S.C. Score of the students who made response 1. In other words, it expresses the deviation, in terms of C.S.C. Scores, of the mean of the response 1 group from the mean of the entire group. The *n*1 column gives for each item the number of students making response 1. And so *da* is the deviation of the mean of the *a* response group from the general mean; *na* gives the

number of cases in the *a* group; and so on. The deviations are in T units and hence in terms of one-tenth of the standard deviation of the C.S.C. Scores of the entire group.

TABLE IV

RESPONSE VALUES IN TERMS OF C.S.C. SCORE, FREQUENCY OF RESPONSES, AND MEAN VALUE OF RESPONSE TYPES OF THE AMERICAN COUNCIL COMPLETION TEST

(Based on Cases 1-137)

Item	*d*1	*n*1	*da*	*na*	*dX*	*nX*	*dy*	*ny*
1	.92	112	–3.84	16	–4.47	9		
2	.10	123	–1.58	9	.53	5		
3	.48	117	–1.97	6	–3.09	14		
4	.47	99	–1.22	27	–3.00	11		
5	.93	112	–4.34	15	–3.87	10		
6	1.07	107	–4.31	19	–2.93	11		
7	.45	99	.58	18	–1.62	20		
8	.40	111	–2.80	3	–1.56	23		
9	.45	99	2.53	12	–3.24	26		
10	.38	81	–3.05	12	.14	44		
11	.76	106	–3.73	6	– .39	25		
12	.04	101	– .04	22	– .15	14		
13	1.16	86	–5.03	9	–1.32	42		
14	.02	129	–8.47	1	.95	7		
15	.75	77	–1.59	16	– .86	44		
16	1.73	31	–1.91	16	– .23	90		
17	.76	96	–3.11	11	–1.27	30		
18	.80	99	–4.47	6	.06	32		
19	.11	90	–2.53	17	1.13	30		
20	.40	47	.04	41	– .41	49		
21	.29	79	1.45	12	.86	46		
22	.59	124	–4.72	4	–5.80	9		
23	2.42	37	– .22	32	–1.21	68		
24	.78	109	–4.14	3	–2.87	25		
25	4.78	4	–4.01	24	.71	109		
26	1.33	65	–1.47	11	–1.64	60	–15.47	1
27	4.61	12	.09	16	– .18	107	–18.47	2
28	2.63	29	–2.23	21	.22	84	–15.80	3
29	1.28	16	.19	68	.10	40	–15.80	3
30	.46	112	–1.16	13	1.10	7	– 8.69	5
31	1.42	83	–5.67	10	.12	34	– 6.47	10
32	.71	113	3.10	7	–3.30	6	– 7.47	11
33	.35	57	–1.63	25	1.69	37	– 2.31	18

TABLE IV (*Continued*)

Item	*d*1	*n*1	*da*	*na*	*dX*	*nX*	*dy*	*ny*
34	.94	39	– .50	29	.42	46	– 1.77	23
38	2.15	39	– .61	21	– .83	50	– 1.07	27
36	1.03	6	2.78	12	– .36	86	– .23	33
37	— .48	60	– .47	21	.83	23	.50	33
38	1.00	60	–3.72	12	– .90	23	.15	42
39	— .47	58	–2.09	13	4.53	5	.60	61
40	3.06	17	14.78	4			– .95	116
Mean (weighted)..	.69		–1.25		– .52		– 1.30	

Table IV reveals the probable need for other than the subjective determination of scoring values. The author assigns a credit of 2 for all responses; to the *a*, the *X*, and the *y* responses is given a uniform value of zero. Still omitting the consideration of the grouping of items, how effective for C.S.C. Score prediction is the author's assignment of scoring values? Apparently the highest effectiveness is not achieved. The truer values within the limits of error for this one group, corresponding to the author's uniform single credit responses range all the way from minus .69 to plus 4.78. The uniform zero values range from plus 14.78 to minus 15.47. In two cases a response credited 1 by the author was, according to the more objective measure, inferior to the responses credited zero by the author. In the case of twelve items, types of so-called "wrong" responses were better than what the author terms "correct" responses. The various types of "wrong" responses differ in value as indicated for each item.

The average (weighted) deviation values at the bottom of the table indicate the differences between the types of responses on the whole. For this test it is in general an indication of about the same ability to omit an item because of lack of time (response *y*) as it is actually to make an incorrect attempt (response *a*) and both these responses are worse than is an omission presumably made after a fruitless consideration of the item problem. The few who are markedly slow (see column *dy*), reaching only as far as Item 31 or 32, are decidedly inferior.

While the same general results would probably be found if the new assigned response values were perfectly reliable, it is

important to note that the values are conditioned by certain chance variations. A discussion of the unreliability of the new response values occurs on pages 29, 44 ff.

What the best method is of scoring true-false or multiple-choice examinations which permit "guessing" has been and still is a matter of contention among test constructors. The purely mathematical treatment of the question resulting in the use of general formulæ like: "$S = R - \frac{W}{(N-1)}$" or "Score equals the number correct minus the number wrong over the number of choices less one," is fallacious, because a host of miscellaneous psychological factors enter to destroy the assumption that guessing in the chance sense has uniformly taken place. The best method of scoring this type of examination can be determined only after much intensive empirical study. Moreover, the best that can be hoped for is that such study will reveal significant types of material for which formulæ, applying generally to a given type, may be constructed. Table V and Figure 2 show the results of the intensive analysis of the items of a true-false academic information test, the Thorndike IIG8.

The form of the table and the symbols employed are similar to those of Table IV. The item numbers run down the column on the extreme left; the *dR* column gives the deviation in C.S.C. Score of the "correct" response group mean from the mean of the whole group; the *nR* column lists the number making the correct response. So with the other symbols: *W* represents "wrong" and *X* and *y* signify types of omissions as in the previous table.

The basis upon which this true-false test is scored by the author is the usual one: 3 being allowed for a correct response, a similar deduction being made for a wrong response, and a zero credit being given for an omission. The test includes materials from the fields of mathematics, biology, psychology, geography, law, music, history, literature, and civics.

Table V and Figure 2 clearly indicate that the test author's scoring does not give to the item responses the values which would be best predictive of the C.S.C. Scores, assuming that the new values approach the true ones. The unweighted mean of the "correct" responses is found to be minus .91T, that of the "wrong" response, minus 1.23T and that of the omission re-

TABLE V

RESPONSE VALUES IN TERMS OF C.S.C. SCORES AND FREQUENCY OF RESPONSES FOR THE SIXTY ITEMS OF THE THORNDIKE TEST IIG8

(Based on Cases 1-100)

Item	*dR*	*nR*	*dW*	*nW*	*dX*	*nX*	*dy*	*ny*
1	—3.31	37	1.64	55	4.01	8		
2	.52	59	3.88	3	—1.12	38		
3	— .09	73	— 4.19	14	5.03	13		
4	—1.24	69	— 0.72	15	6.00	16		
5	.41	83	— 3.16	14	3.88	3		
6	— .98	51	— .69	30	3.72	19		
7	1.29	81	— 5.66	11	—5.25	8		
8	—1.67	53	— 1.52	15	3.31	32		
9	.28	52	1.38	22	—2.70	26		
10	.59	87	.13	4	—5.79	9		
11	1.20	63	— 6.52	5	—1.34	32		
12	— .49	32	— 1.04	13	.53	55		
13	1.13	24	— 1.12	34	.26	42		
14	—2.37	20	.05	23	.81	57		
15	.97	26	— 4.12	6	.74	68		
16	—7.12	9	.88	2	.70	89		
17	— .23	57	— 1.30	11	.85	32		
18	.22	41	— 3.32	10	.49	49		
19	.37	57	— 3.49	19	1.88	24		
20	—1.95	12	— 2.01	38	2.00	50		
21	.71	41	— 2.05	29	—1.25	30		
22	—3.16	27	— 3.12	7	1.62	66		
23	1.38	12	— .36	84	3.38	4		
24	— .38	38	— .87	44	2.94	18		
25	1.32	32	— 1.88	55	4.73	13		
26	1.65	22	— 4.23	18	.66	60		
27	— .13	73	— 9.62	2	1.16	25		
28	— .20	92	13.88	1	.74	7		
29	—4.21	32	.45	30	3.20	38		
30	— .15	32	— 2.04	38	2.75	30		
31	—2.35	17	— 6.18	17	2.20	66		
32	—5.42	10	— 2.72	15	1.27	75		
33	.48	5	— 3.38	23	1.05	72		
34	—2.79	21	— 1.37	31	2.17	48		
35	—1.75	32	.23	23	1.12	45		
36	— .08	94	—18.12	2	10.88	4		
37	.39	45	— 1.81	16	.29	39		
38	— .51	73	3.38	2	1.24	25		
39	—3.54	36	.45	7	2.17	57		
40	.94	70	— 2.51	23	—1.26	7		

TABLE V (*Continued*)

Item	*dR*	*nR*	*dW*	*nW*	*dX*	*nX*	*dy*	*ny*
41	— .83	81	5.32	9	1.88	10		
42	—2.53	34	— .86	39	4.44	27		
43	.26	29	—2.26	47	4.15	24		
44	.07	95	....	..	—1.08	5		
45	— .14	45	— .60	33	.79	21	—32.12	1
46	—4.34	33	4.71	12	2.20	54	—32.12	1
47	—5.65	19	—1.62	6	2.02	74	—32.12	1
48	— .03	88	2.50	8	5.33	3	—32.12	1
49	1.52	55	1.08	25	4.29	19	—32.12	1
50	— .42	70	.13	16	4.49	13	—32.12	1
51	.13	52	—1.17	19	1.65	28	—32.12	1
52	— .67	75	—2.45	6	4.38	18	—32.12	1
53	—1.12	27	.14	19	2.07	51	—16.12	3
54	.67	52	—1.27	20	1.56	25	—16.12	3
55	—1.34	24	—3.06	19	2.57	54	—16.12	3
56	.30	83	—1.72	10	10.13	4	—16.12	3
57	.39	91	3.38	4	— .12	2	—16.12	3
58	—1.64	21	1.74	14	—1.60	27	1.70	38
59	—4.87	12	— .12	10	— .55	27	2.08	51
60	—1.85	33	—2.43	13	...	..	1.71	54

sponse, plus 1.53T. To confess a lack of knowledge proved to be predictive of better college work than did making either a correct or an incorrect response. The logically deduced scoring does not provide for this type of situation; the empirically determined key does, within the limits of chance variations discussed on page 29 below, make such provision, whether for better or worse. In twenty-seven instances the "wrong" response receives a higher value in terms of college achievement than the "correct" response. These reversals may be due partly to the unreliability of the C.S.C. Scores, to the operations of chance, to a low correlation between the item responses and the C.S.C. Scores or to some unforeseen ambiguity in the item. But the persistently high value of the omission response is a practically certain indication that chance error is not the only significant factor operating. It is safe to say that there is an underlying value to each response in each item which would be roughly approximated for a large group studied by such results as those of Table V.

With this test, as with the American Council Completion, the

few who are especially slow (see column marked "*dy*") are markedly inferior in C.S.C. Score. The last three *dy* values are of doubtful significance because, the last questions being highly technical, many students may have omitted all three, although they actually had time to consider them.

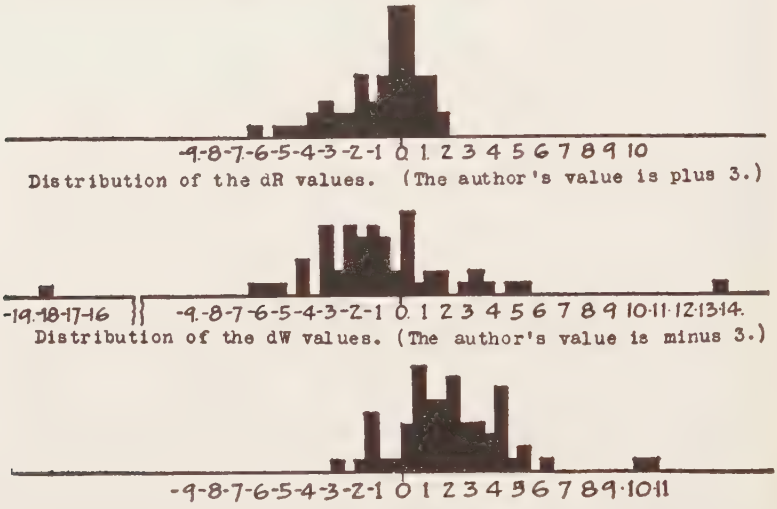

FIGURE 2. THE dR, dW, AND dX VALUES OF TABLE V REPRESENTED GRAPHICALLY

The vertical axis indicates the number of items falling within the half-unit T deviation intervals denoted along the horizontal axis.

The Thorndike IIG3 is a three-minute, ten-item picture completion test. Certain miscellaneous points are revealed with the help of the objective analysis of response values, tabulated in the usual manner in Table VI. Responses 2 and 1 are respectively totally correct and partially correct completions. Response *m* is a completion which, while not creditable according to the author's scoring key, seemed to the scorer to be worth crediting. Responses *a*, *X*, and *y* have the same connotation as in Table IV.

The results with Item 1, for example, are noteworthy. The item is very easy, only four students having failed to make the correct response, and yet the four who did not attempt a re-

sponse were markedly above the average of the group in C.S.C. Score. Evidently, some element in the item situation was present to several superior students which is not apparent and which destroys whatever value the item might have had unless better scoring values are assigned.

TABLE VI

RESPONSE VALUES IN TERMS OF C.S.C. SCORES AND FREQUENCY OF RESPONSES FOR THE ITEMS OF THE THORNDIKE TEST IIG3

(Based on Cases 1-100)

Item	*d*2	*n*2	*d*1	*n*1	*dm*	*nm*	*da*	*na*	*dX*	*nX*	*dy*	*ny*
1 ..	— .55	96							13.13	4		
2 ..	.29	84					— 1.43	16				
3 ..	.00	100										
4 ..	— .08	**99**									16.12	1
5 ..	2.29	49	—2.13	45	4.88	1	— 18.12	1	8.88	1	11.50	2
6 ..	6.59	**7**			5.38	2	— 3.57	11	— .03	74	— 2.62	6
7 ..	2.32	11	— .47	23	1.88	1	— 1.92	15	.90	43	— 2.36	7
8 ..	.01	60					2.74	7	— .52	25	— 1.97	8
9 ..	— .29	6	.64	72			4.12	3	6.88	10	— 1.68	9
10 ..	— .07	41			— 3.12	3	— 2.72	23			.64	33

The separate tabulation of doubtful responses offers a means of corrective modification of a tentative scoring key. Thus in Items 5, 6, and 7, the doubtful responses (which are available on the original test blanks) might well be included with the correct ones, while in Item 10 the doubtful response appears to be no better than the incorrect responses.

The values of the omissions not due to lack of time again compare well with the other types of response. The *dy* column reveals some slight additional data on the elusive problem of test speed as an indication of college success ability.

The Brown Test 1 consists of twenty items of the following type:

The.....of history has no.....unless it helps to .....the present.

In contrast with the American Council Completion Test, the Brown Completion is characterized by a greater number of omissions to the sentence, the use of more common words, no indication of the number of letters contained in each omitted

word, and an emphasis upon the syntactical or grammatical structure of language rather than on diction. The sentence and not each omission is regarded as the item unit.

Table VII has the same general form as Tables IV, V, and VI. The author distinguishes between two qualities of "correct" re-

TABLE VII

RESPONSE VALUES IN TERMS OF C.S.C. SCORES AND FREQUENCY OF RESPONSES FOR THE ITEMS OF THE BROWN TEST 1

(Based on Cases 1-100)

	Item	*d*2	*n*2	*d*1	*n*1	*dA*	*nA*
E	1	1.80	45	— 2.71	18	— .87	37
	2	.28	77	1.57	6	—1.84	17
	3	—6.10	2	.60	41	— .21	57
	4	— .25	98	16.40	1	16.40	1
	5	3.30	19	— 3.46	7	— .52	74
	6	1.17	47	—13.10	2	— .54	51
	7	5.73	6	.45	19	— .57	75
	8	— .82	14	1.32	13	— .07	73
	9	.57	23	1.17	40	—1.63	37
	10	—3.10	18	3.98	12	— .03	70
F	1	.33	89	— 9.60	1	—2.00	10
	2	.40	50	— .47	16	— .36	34
	3	.47	44	1.15	4	— .38	52
	4	— .73	31	— 7.35	6	.84	63
	5	.88	54	— 1.30	23	—1.10	20
	6	2.84	27	.60	5	—1.17	68
	7	11.90	2	4.08	19	—1.28	79
	8	— .71	27	.15	12	.29	61
	9	17.00	1	— 8.60	2	.00	97
	10	2.40	7	1.07	3	— .29	90

sponses. Credits of 2 and 1 are allotted according to this distinction. The response symbols 2 and 1 retain the significance assigned by the author. Any wrong response (including omissions, since they were few) is included under *A*. The response values in terms of criterion scores of the Brown Completion items show approximately the same degree of variability for each response type as do those of the other tests discussed above.

The results of the analysis of the Brown Test 3, a twenty-item opposites test, are indicated in Table VIII. The items are of the form:

Joy: grief, sad, sorry, ride, discomfiture, happiness, scowl, enjoy.[1]

The student is asked to underline the word which is the exact opposite of the first word. In the table, response 1 designates the correct response, response *A* any other response. The

TABLE VIII

RESPONSE VALUES IN TERMS OF C.S.C. SCORES AND FREQUENCY OF RESPONSES FOR THE ITEMS OF THE BROWN TEST 3

(Based on Cases 1-100)

	Item	*d*1	*n*1	*dA*	*n*A
E	1	0.00	100	0.00	0
	2	.12	95	2.20	5
	3	.20	73	.53	27
	4	— .16	95	3.00	5
	5	.08	97	— 2.60	3
	6	.10	99	— 9.60	1
	7	.33	87	— 2.21	13
	8	.85	77	— 2.86	23
	9	— .07	75	.20	25
	10	1.65	36	— .93	64
F	1	0.00	100	0.00	0
	2	0.00	100	0.00	0
	3	— .10	90	.90	10
	4	— .06	99	7.60	1
	5	— .13	97	5.27	3
	6	— .01	94	.07	6
	7	.22	93	— 2.89	7
	8	— .01	63	.02	37
	9	— .80	40	.53	60
	10	— .48	54	.44	46

form of the table is similar to that of Table VI. Because of the small degree of difficulty of most of the items, little diversity of response values is evident. Within a diminished range the general conclusions of the four preceding tables are substantiated.

The Roback Tests 1 and 7 were selected for item analysis

[1] Selected from the practice form of the test.

especially because of the highly subjective nature of their scoring keys. Test 1 calls for writing the general term which expresses the general class under which each of a series of words, such as man, lion, amœba and ant, is included. Test 7 requires

TABLE IX

RESPONSE VALUES IN TERMS OF C.S.C. SCORES AND FREQUENCY OF RESPONSES FOR THE ITEMS OF THE ROBACK TEST 1

(Based on Cases 1-100)

Item	*dr*	*nr*	*d1*	*n1*	*dm*	*nm*	*da*	*na*	*dX*	*nX*	*dy*	*ny*
1 ..	1.29	23					— .11	34	— 2.39	11		
2 ..	—4.05	16					.67	38	2.08	14		
3 ..	.94	32					— .12	35	32.51	1		
4 ..	— .40	46					—2.99	12	5.41	10		
5 ..			—1.18	29			.90	33	.67	6		
6 ..	— .48	61	5.51	1			3.51	5	7.51	1		
7 ..	— .71	50					1.39	17	—11.49	1		
8 ..	.19	50					— .73	17	2.51	1		
9 ..	— .81	43			5.01	2	— .06	14	2.84	9		
10 ..	—1.93	41					—1.51	12	.51	15		
11 ..	— .80	42					— .49	18	5.26	8		
12 ..	—1.43	35					.06	22	4.46	11		
13 ..	.01	18					— .18	35	2.18	15		
14 ..	— .33	51					1.45	16	— 6.49	1		
15 ..	.54	39			—1.09	10	.51	16	7.84	3		
16 ..	.03	44					—1.00	20	4.76	4		
17 ..	.56	43					.22	7	— .72	17	9.51	1
18 ..	— .19	50					—1.42	14	6.51	3	9.51	1
19 ..	—1.55	16	— .25	42			2.84	9			9.51	1
20 ..	.14	32					—2.25	21	3.01	14	9.51	1
21 ..	.03	26	— .72	13			.47	26	— 5.99	2	9.51	1
22 ..	—1.35	17	2.01	2			— .72	38	3.71	10	9.51	1
23 ..	—1.33	25	13.51	1			—2.79	29	3.76	12	9.51	1
24 ..	—1.18	39					.51	21	3.55	7	9.51	1
25 ..	— .68	16					— .28	43	.80	7	8.51	2
26 ..	— .17	19					—1.92	21	.80	26	8.51	2
27 ..	—2.40	11			4.76	4	.02	39	— .90	12	8.51	2
28 ..	.57	32	4.88	8			—1.76	15	— 4.49	10	6.51	3
29 ..	— .73	41	1.51	1			.01	20	3.01	2	5.51	4
30 ..	3.51	11			—2.72	13	—1.22	34			2.81	10

the subject to write in the word that is the exact opposite of each printed word.

The types of responses for which C.S.C. Score values are presented in Tables IX and X are as follows: correct according to the author; partially correct according to the author; doubtful according to the scorer; incorrect word; omission with later attempt in the same test; and omission with no later attempt. The symbols are as before: 2, 1, *m*, *a*, *X*, and y in the order mentioned.

TABLE X

RESPONSE VALUES IN TERMS OF C.S.C. SCORES AND FREQUENCIES OF RESPONSES FOR THE ITEMS OF THE ROBACK TEST 7

(Based on Cases 1-68)

Item	*d*2	*n*2	*d*1	*n*1	*dm*	*nm*	*da*	*na*	*dX*	*nX*	*dy*	*ny*
1 ..	— .36	47			1.41	19	—4.99	2				
2 ..	.20	64			—3.24	4						
3 ..	1.01	2			—3.37	17	1.31	40	1.18	9		
4 ..	—1.07	19	—2.43	18	3.18	6	1.15	22	6.51	3		
5 ..	—1.32	30			2.06	29	—3.49	8	7.51	1		
6 ..	—3.78	7			1.43	12	— .79	44	8.80	5		
7 ..	—1.71	9	1.34	6	1.93	7	.94	35	—3.67	11		
8 ..	1.26	48			—3.49	3	—2.70	14	—4.16	3		
9 ..	—3.92	7	—6.49	1	— .21	27	— .67	22	4.87	11		
10 ..	— .05	57			—2.49	3	4.18	6	—7.49	2		
11 ..	—2.39	21	2.81	13	.03	25	.67	6	1.84	3		
12 ..	—4.49	4	— .49	5	1.50	48	—4.27	9	—7.49	2		
13 ..	.66	26			1.51	1	.74	35	—6.16	6		
14 ..	—3.12	8	1.53	41	—1.18	13	—1.01	6	—4.32	6		
15 ..	.14	54			—5.09	5	1.26	8	7.51	1		
16 ..	3.88	8	2.01	4	.51	21	— .94	36	.18	9		
17 ..	—2.64	13	3.81	10	1.42	11	— .49	38	.76	4		
18 ..	— .86	58			—2.49	3	19.51	3	.76	4		
19 ..	— .36	46			— .90	16	1.24	11	7.51	1		
20 ..	— .06	11			— .19	33	— .04	20	— .16	3	7.51	1
21 ..	—6.43	13			— .94	20	1.89	24	3.81	10	7.51	1
22 ..	—6.99	2			— .22	56	1.34	6	3.51	3	7.51	1
23 ..	— .10	54			— .05	5	3.01	4	—1.99	4	7.51	1
24 ..	— .58	11	3.19	15	—1.49	13	—1.58	23	— .99	4	8.51	2
25 ..	—2.34	7	—2.06	14	1.97	26	— .12	16			— .78	5

Several outstanding features of Table IX are:

The number of 1 and *m* responses is so few as to make a separate treatment of these types infeasible.

The C.S.C. Score values of the 2, or "correct" responses are no higher in general than those of the *a*, or "incorrect" responses.

The omission responses, *X* and *y*, are in general higher in C.S.C. Score values than any other responses.

The one student who completed only slightly more than half the test was markedly superior in college ability, as were the three others who were able apparently to get no further than the twenty-eighth item.

Table X reveals somewhat similar results:

The "doubtful" response values are in general superior to the "correct" response values.

The *a* response values are on the average about equal to the *m* response values and slightly below the *X* response values.

The two students who were unable to reach as far as the next to the last item were decidedly superior in C.S.C. Scores.

The relatively low indication of a definite trend in the C.S.C. Score values of the various response types reflects in the cases of both tests the low correlation between the C.S.C. Scores and the scores on these tests. Individual items, however, show evidence of worth as indicated by Tables IX and X and by the treatment in Chapter IV of some of the items.

EMPIRICAL COMPARISON OF THE OLD WITH THE NEW SCORING METHODS

That there exist wide and varied discrepancies between the usual and the new scoring values is clear. Disagreement, however, proves the case for neither method of scoring. Only an empirical comparison of the two methods will yield any definite evaluation. Such a comparison has been attempted, but it will probably prove helpful in the interpretation of the comparative results to consider first the more significant sources of error associated with each of the procedures.

The evaluation of the item responses on the usual basis involves in the main two inaccuracies: that of assigning a single value to responses that actually merit a wide range of values,

and that of omitting to take into account unforeseen psychologically effective elements in the item or test situation.

Although planned to minimize these difficulties, the new scoring values are not entirely free from error. First, the values, being dependent upon the criterion scores, are subject to the unreliability and the possible invalidity of these scores. Second, it is at times impossible, at times infeasible, to analyze the possible responses into all of the significant groups. Third, values determined with one sample of a population are subject to errors of sampling or to errors due to changed conditions or varied selection, when applied to a second group.

Certain other sources of error are common to both procedures. It is not now clear which procedure suffers most through them. Thus, whenever an item involves guessing, the scoring values tend to become less significant. Moreover, neither the old scoring procedure nor the new technique as employed to this point, takes into consideration the intercorrelations among the items of a test. The question of unreliability of the new scoring values is again discussed on pages 44 ff in connection with a consideration of the unreliability of the new item coefficient.

To determine the relative effectiveness of the errors associated with each of the scoring procedures, and consequently the relative worth of these procedures, the following steps were taken:

1. Determining by means of correlation with the C.S.C. Scores the validity of a given subtest, when scored according to the author's method.
2. Rescoring the items of the subtest by the new method.
3. Determining the validity of the subtest when scored by the new method.
4. Comparing the old method validity coefficient with the new.
5. Interpreting the relative size of the coefficients in the light of associated data.

To avoid an unfair comparison, it is necessary to determine validity coefficients for both the new and the old scoring with a group other than that used in the computation of the scoring values. Unless definitely otherwise indicated these validity coefficients are worked out with a different group from the one used in determining the scoring values. To facilitate the exposition of the results, the term "basal" is applied to the

group with which the scoring values were computed, while the word "trial" refers to the second group upon which the comparative validity coefficients were computed.

Part of the evidence as to the value of the new scoring method is presented in Table XI. In that table, column I gives the names of the subtests studied, column II, the number of items

TABLE XI

DATA ON THE EMPIRICAL EVALUATION OF THE OLD AND NEW SCORING METHODS

I Subtest	II No. of Items	III Old Scor. Val. Coef. All cases	IV No. of Cases	V Old Scor. Basal Group	VI No. of Cases	VII Old Scor. Val. Coef. Trial Group	VIII New Scor. Val. Coef. Trial Group	IX No. of Cases
American Council 1 Completion	40	.252	137	.161	68	.340	.345	69
Thorndike IIG8 .... Academic Information	60	.099	175	.010	100	.206	—.011	75
Thorndike IIG3 ... Picture Completion	10	.042	175	.061	100	.021	.059	75
Brown 1 ..... Completion	20	.230	248	.230	100	.177	.146	100
Brown 3 ..... Opposites	20	.250	248	.162	100	.247	.220	100
Roback 1 ..... Abstraction	30	.060	137	—.192	68	.258	—.148	69

contained in each test, column III, the coefficient of correlation between the author's scores and the C.S.C. Scores, employing all the cases as indicated by the number in column IV. Column V contains the correlation coefficients between the author's scores and the criterion scores with only those cases employed in determining the new scoring values. The number of such cases is given in column VI. Columns VII and VIII give the significant validity coefficients according to the old and the new method of scoring, respectively.

Table XII indicates the difference between the size of the validity coefficients with scores based on the old scoring method on the one hand and with those based on the new technique on the other. The table also indicates the reliability of these differences. Thus the results with two tests point slightly in

favor of the new technique, two slightly against it, and two markedly against it. The results with these last two tests alone are sufficiently reliable to show that a true difference quite certainly exists. Thus far, then, with the material employed in this study the use of the new scoring technique has failed to produce any significant improvements in the terms of the size of the validity coefficients, and in two cases it has apparently caused a marked deterioration.

TABLE XII

DIFFERENCES BETWEEN THE OLD AND NEW VALIDITY COEFFICIENTS AND THE RELIABILITY OF THE DIFFERENCES

Subtest	Difference between Old and New Validity Coefficients		P.E. of the Difference	Diff. ÷ P.E. Diff.	Chances in 100 of True Difference Greater than Zero
	In Favor of Old	In Favor of New			
American Council 1 .... Completion		.005	.102	.049	51
Thorndike IIG8 ..... Academic Information	.217		.107	2.028	94
Thorndike IIG3 ..... Picture Completion		.038	.106	.356	59
Brown 1 ..... Completion	.031		.093	.333	59
Brown 3 ..... Opposites	.027		.090	.300	58
Roback 1 ..... Abstraction	.406		.110	3.691	99

If it were possible to locate the cause of the results with each test, it might be possible to state at least under what circumstances the new technique might prove of value.

The results reported in Tables XI and XII in addition to other data give some tentative suggestions as regards the solution to this problem. Following are certain outstanding conclusions relative to the attempt to discover the factors associated with the effectiveness of the new scoring method:

1. The number of items contained in a test seems to have no significant association with the effectiveness of the new scoring technique.

2. Excepting for the fact that no test having a high old scor-

ing validity as computed with all the cases[1] showed a marked deterioration when the new method was applied, there seems to be no clear evidence of the association of this validity measure and the effectiveness of the new method.

3. The number of cases with which the new scoring values were determined did not vary sufficiently to give any indication of the effect of this factor upon the value of the new procedure.

4. The effectiveness of the new scoring technique does seem to depend upon the relative size and direction of the old scoring validity with the basal group as compared with the old scoring validity with the trial group. The two serious reversals indicated in Tables XI and XII seem to be due to the fact that the basal validity coefficients are markedly below the trial validity coefficients.

Substantiating evidence of this relationship was obtained in the instance of the American Council Completion Test by computing new scoring values with the trial or second group, rescoring the test responses of the basal or first group on the basis of these values and then making the usual comparison of the validity coefficients based on the two methods. The results were:

Old Scoring Validity with Cases on which New Scoring Values Were Determined	.340 ± .072
Old Scoring Validity with a Second Group	.161 ± .079
New Scoring Validity with a Second Group	.252 ± .077

Associated with the use of the new technique is an improvement of .091 in the validity coefficient.

Only in the instance of the Brown Completion Test did the new scoring validity coefficient as compared with the old scoring validity coefficient with the trial group move distinctly opposite to the old scoring validity coefficient with the basal group. The dissimilarity between the basal and the trial groups appears to be a significant determinant of the relative success or failure of the new scoring method. This dissimilarity is roughly indicated by the difference between the old scoring validity coefficients when computed with the basal group on the one hand and with the trial group on the other. The following coefficients indicate the effectiveness of the new technique with the American Council Completion Test when the trial group is identical with the basal

[1] See Table XI, column III.

group (and incidentally when the size of the basal group is increased):

Old Scoring Validity Coefficient (137 cases)	.252 ± .054
New Scoring Validity Coefficient (137 cases)	.511 ± .043

Since approximately the highest possible coefficient (within the limits of chance variation) is .559 (due to the unreliability of the criterion scores), the amount of this difference is remarkable. However, since practically no two groups ever reach even approximate identity, the improvement, practically considered, offers little encouragement.

The new scoring technique seems to have failed essentially because of the lowness of the old scoring validity with the basal group and because of the dissimilarity between the basal and the trial group. If the new method is to be of practical utility, it must prove its worth with groups as dissimilar as those here employed, but it ought to receive a trial with tests showing higher validity coefficients than those used here. This implies experimentation in a field where criterion scores that are highly reliable, can be found.

TENTATIVE TRIAL OF MODIFICATIONS OF THE METHOD OF DETERMINING NEW SCORING VALUES

Notwithstanding the fact that the new scoring method seems thus far to have proved unsuccessful largely because of the unreliability of the criterion scores, it might prove serviceable for future work, to note, even with the use of the same scores, the effects of modified scoring procedures. As material for this preliminary study the ten best items of the Thorndike True-False Academic Information Test, as chosen by the modified item coefficient described on page 50, were employed. The distribution of the C.S.C. Scores of those making each response for each of the items is presented in Table XVIII, together with the frequency of each response and its mean C.S.C. Score value.

In the table, the C.S.C. Score value in terms of the deviation from the approximate median is indicated vertically at the left. The frequency of the scores for the entire group of 100 is listed in the next column. The remaining columns give the frequency of the C.S.C. Score for each response group. The total frequencies and the means of the various groups are indicated at the foot of each column. The tabulation is curtailed.

A glance at the distributions of Table XIII will clarify the problem involved in determining an objective value for each response and in addition will suggest possible measures that might be used to represent the response value. The mean C.S.C. Score of each response distribution ought, by general theory, to

TABLE XIII

DISTRIBUTION OF C.S.C. SCORES FOR THE RESPONSE GROUPS WITH THREE ITEMS OF THE THORNDIKE IIG8 TEST

C.S.C. Score	Score F	Item 26			Item 3			Item 55			
		fR	*fW*	*fX*	*fR*	*fW*	*fX*	*fR*	*fW*	*fX*	*fy*
19	4	1	0	3	2	0	2	1	0	3	0
16	5	3	0	2	4	0	1	1	1	3	0
14	1	0	0	1	1	0	0	0	0	1	0
13	1	0	0	1	0	0	1	0	0	1	0
12	2	0	0	2	2	0	0	0	0	2	0
10	5	1	1	3	3	2	0	2	0	3	0
9	3	0	1	2	2	0	1	0	1	2	0
7	6	1	1	4	4	1	1	1	1	4	0
5	5	4	1	0	4	0	1	2	1	2	0
4	8	0	1	7	6	1	1	1	0	7	0
2	6	1	1	4	5	1	0	3	2	1	0
1	2	1	0	1	1	1	0	0	1	1	0
0	5	2	0	3	5	0	0	1	0	3	1
—2	5	2	0	3	5	0	0	0	4	1	0
—3	5	3	0	2	3	1	1	0	0	5	0
Frequency.....	100	22	18	60	73	14	13	24	19	54	3
Mean		1.7	—4.2	.7	—.1	—4.2	5.	—1.3	—3.1	2.6	—16.1

represent best the desired value. It is conceivable that in this special type of situation, some other measures, like the median of the Q_3, might serve better. These two measures are usually less reliable than the mean, but eliminate, in common, the emphasized effect of the extreme cases. The Q_3, in addition, tends to disregard the distribution of the inferior cases, who may be expected, in general, to fall very often by chance into the response groups. A third measure permits those having a C.S.C. Score above the median criterion score to judge, as it were, the order and the degree of value of the various responses, by assigning to each response a value proportional to the number of the superior group making the given response.

The various measures of response values described above were

computed for the ten items. In the case of practically all the items, the difference between the lowest and the highest was reduced to five units. Scores on the ten-item set were determined according to the various proposed rescoring methods and validity coefficients, correlating these scores with the C.S.C. Scores, were computed. The validity coefficients of correlation were:

Author's Scoring	$.210 \pm .075$
Mean as Response Value	$.112 \pm .077$
Q_3 as Response Value	$.139 \pm .077$
Median as Response Value	$.173 \pm .076$
Proportion of Superior as Response Value	$.085 \pm .077$

The same 100 cases was used; the ten items were the same; the only variant was the method of scoring. In the case of the proportional method of assigning scoring values, two items were practically undifferentiating, and hence the score was determined with only eight items. The coefficients are computed with a group other than that employed in determining the various new scoring values.

That reducing the difference between the highest and lowest response value uniformly to five for each item was an improvement over the varied differentiation used heretofore, is indicated by the following pairs of validity coefficients of correlation computed with new scores based on the ten items employed above.

Validity Coefficients Employing the Mean with Unlimited Differentiation	$.097 \pm .077$
Validity Coefficients Employing the Mean with Limited Differentiation	$.112 \pm .077$
Validity Coefficients Employing the Q_3 with Unlimited Differentiation	$.108 \pm .077$
Validity Coefficients Employing the Q_3 with Limited Differentiation	$.139 \pm .077$
Validity Coefficients Employing the Median with Unlimited Differentiation	$.139 \pm .077$
Validity Coefficients Employing the Median with Limited Differentiation	$.173 \pm .076$

The partial evidence here presented points, then, to the use of the median as the best of the response value measures and to the approach at uniform weighting of items by limiting the distance between the lowest and the highest response value to a set number of units. A third tentative suggestion is to employ

percentile rank units in the measures of the criterion trait. This would minimize the effect of the extreme cases and yet permit the calculation of the mean. This suggestion is unsupported by any empirical evidence as to its worth. The new methods tried with the ten items have failed to excel the old method, but point to possible added lines of approach to the problem of the improvement of the scoring of tests.

CHAPTER IV

THE PROBLEM OF THE CHOICE OF THE BEST ITEMS

CHOICE OF THE ITEM COEFFICIENT

The introductory chapter presents in rough outline the hypothetical solution to the problem of the choice of the best items for a subtest or an examination. The theoretically perfect procedure would involve the determination of the validity coefficients of all available test items, the computation of the intercorrelation of each possible pair of items for the entire examination and the building of regression formulæ (with accompanying regression weights) for the thousands of possible combinations of items. The use of intercorrelations of all pairs of items is obviously outside the pale of practicality, and hence a substitute must be found. A feasible substitute is commonly used by test constructors; namely, to regard all the items within a given subtest score as the element in determining regression weights. Another feasible substitute involving a modified form of the principles underlying the intercorrelations of items is developed later on page 49. But in the main, items must be selected on the basis of their validity, that is, on the basis of the effectiveness with which they predict significant criterion scores.

It is therefore essential that test builders be equipped with a sound method of computing a coefficient to represent the validity of an item. In the search for a coefficient the first thought is to attempt to apply the various coefficients of association or correlation used in usual statistical treatments. Vincent, after considering various possible measures, decided to use the method of overlapping.

The most obvious deficiency of that method is that it may be applied to only a twofold classification of item responses. The distinct need was felt for a coefficient that would apply to items yielding three or more categories of responses as well as to those yielding only two.

McCall has invented a coefficient which eliminates this deficiency. It is an outgrowth of the principle of assigning scoring values according to the mean of the criterion scores of the group making a certain response. The size of the coefficient is dependent in part upon the distances between each pair of response means. The second determining factor is the product of the frequencies of the responses taken in every possible combination of two. These two factors are combined to give the numerator of the item validity formula:

$$C = \frac{(M_1 - M_2)(N_1 \times N_2) + (M_1 - M_3)(N_1 \times N_3) + (M_2 - M_3)(N_2 \times N_3) \ldots\ldots \text{etc.}}{\text{S.D.} \times N^2}$$

where:

C is the coefficient;

M_1, M_2, M_3, etc. are in their order of size from highest to lowest, the means in terms of criterion score of the groups making respectively response 1, 2, 3, etc. for the given item;

N_1, N_2, N_3, etc. are the frequencies of the respective responses, 1, 2, 3, etc.;

S.D. dist. is the standard deviation of the entire group of which the response groups are component parts;

N_2 is the square of the frequency of the entire group.

Letting d_{12} represent $M_1 - M_2$, etc., the formula may be more conveniently written:

$$C = \frac{d_{12}n_1n_2 + d_{13}n_1n_3 + d_{23}n_2n_3 \ldots}{\sigma N^2}$$

An empirical study of various possible item coefficients including the above is being made by Miss H. Barthelmess. At the time it became necessary to select an item coefficient for the present study, Miss Barthelmess very kindly gave her advice on the basis of whatever data were then available in her study. The data pointed to the superiority of the coefficient described above and hence it was selected for use.

The original purpose of this phase of the study was to discover the relation between item validity and certain other characteristics of items, such as consistency with other items, difficulty, form, and so on. Before any such relations can be safely reported to exist or not to exist it is essential that the effective-

ness of the item coefficient be examined. It is furthermore necessary to note any spuriously associated elements in the measures to be related, such as the validity coefficient and the difficulty measure. Consequently, a consideration of the characteristics of the coefficient and an empirical evaluation of its effectiveness are essential.

CHARACTERISTICS OF THE ITEM COEFFICIENT

1. Bliss has shown that if N^2 is used in the denominator, coefficients with groups of varying sizes become comparable. The algebraic proof is simple:

$$C = \frac{d_{12}n_1n_2 + d_{13}n_1n_3 + d_{23}n_2n_3 \ldots\ldots}{\sigma\ N^2}$$

Assuming: first, that the deviations between pairs of response values are constant; second, that the standard deviation of the entire distribution remains identical with the increased number; and third, that the response group frequencies remain proportionally similar; then, increasing the size of N by a results in the following:

$$C = \frac{d_{12}(an_1)(an_2) + d_{13}(an_1)(an_3) + d_{23}(an_2)(an_3) \ldots\ldots}{\sigma\ (aN)^2}$$

$$= \frac{d_{12}n_1n_2a^2 + d_{13}n_1n_3a^2 + d_{23}n_2n_3a^2 \ldots\ldots}{\sigma\ N^2a^2}$$

Dividing both numerator and denominator by a^2, the original formula ensues.

However, the first of the above assumptions may involve a constant error in that chance errors tend to increase the size of the coefficient more when N is small than when it is large.

2. The division by the measure of variability makes comparable coefficients computed with groups of varying degrees of dispersion. Logically, the S.D. appears to be the best measure to employ because of its reliability and because it weights heavily the extreme cases, thus equalizing the effect of such cases upon the deviations of the numerator. However, mathematical proof on this point is wanting.

3. Other things being equal, the item coefficient formula weights heavily the equal division into response groups. Thus, assuming a deviation of one between the values of Response 1 and Response 2, the numerator of the coefficient would vary as

follows for a total group of ten cases, n_1 and n_2 being the frequencies of the respective response groups:

n^1	n^2	numerator
0	10	0
1	9	9
2	8	16
3	**7**	**21**
4	6	24
5	5	25
6	4	24
7	3	21
8	2	16
9	1	9
10	0	0

The denominator remains the same in each case and hence the coefficient varies in proportion with the above numerators. Items containing responses that are neither too easy nor too difficult are consequently favored, notably when only two types of responses are recorded. This characteristic may work perniciously in the case of items allowing for guessing, like true-false items. The guessing error tends to equalize the size of the response groups and slight erratic deviations between these subgroups are consequently emphasized. Moreover, in attempting to determine the relation between item goodness and the difficulty of the item, as will be shown later, this characteristic of the coefficient tends to destroy the validity of the results found.

4. The size of the item coefficient is dependent upon the adequateness of the analysis of the significant types of responses. Thus, if two significantly different responses having consequent differing true response values are first treated as distinct and then are included in the same category, the resultant coefficient will vary according as the response values were combined or treated separately. In comparing items, then, a similar degree of carefulness of response analysis, consistent with feasibility of scoring, ought to maintain for each item studied. Standardization of principles underlying item analysis would be helpful.

Disregarding effects due to chance, the size of the coefficient is not increased merely by increasing the number of response categories unless a more valid distinction among responses actually is made. Thus, if a response type has a C.S.C. Score value of 58 and a frequency of 20, for example, to separate the responses

into two types with frequencies of say 15 and 5, or 10 and 10,—types having no true distinction,—will yield for each the same value, namely 58, provided the effects of chance variations are omitted. The chance variations associated with the formation of types not truly significant tend to increase the size of the item coefficient.

5. In the item coefficient formula the difference between the response values play an important part. Should these differences be expressed in whole numbers or in finer units? What is the effect on the size of the coefficient of the degree of fineness used in determining differences? Theoretically, the coefficient ought in general to increase in size with the refinement of units because of the greater differentiation between some response values which otherwise might have been regarded as identical. To discover the actual effects, the coefficients of the forty items of the American Council Completion Test were computed first, employing whole number differences between response values and second, using two place decimal units. The coefficients by the first method were correlated with those by the second. The coefficient of correlation was found to be .909. The mean of the crude unit difference coefficients is .065, that of the other is .070. The indications are, then, that while the relative positions of the coefficients remain very nearly the same regardless of the refinement of units, this refinement tends to increase the coefficient.

6. The item coefficient formula calls for the multiplication of the difference between the pairs of response means by the product of the frequencies of the respective responses. It is significant to note that the M values and the n values are not unrelated. The contingency diagram of Figure 3 expresses this association, the M values having been transmuted into deviations from the general mean of the C.S.C. Scores.

The vertical axis represents the difference between a particular response value in terms of the C.S.C. Score and the mean C.S.C. Score of the entire group, while the horizontal axis indicates the frequency of the response. The data on which the contingency diagram is built may be found in Table IV on page 17. At the lower left-hand corner of the diagram of Figure 3 is indicated, for example, that of the response values between 0 and 105, eight were associated with frequencies of from 1 through 21, and so on.

The line of the maximum deviation for various frequencies indicates the values that items showing perfect differentiation would

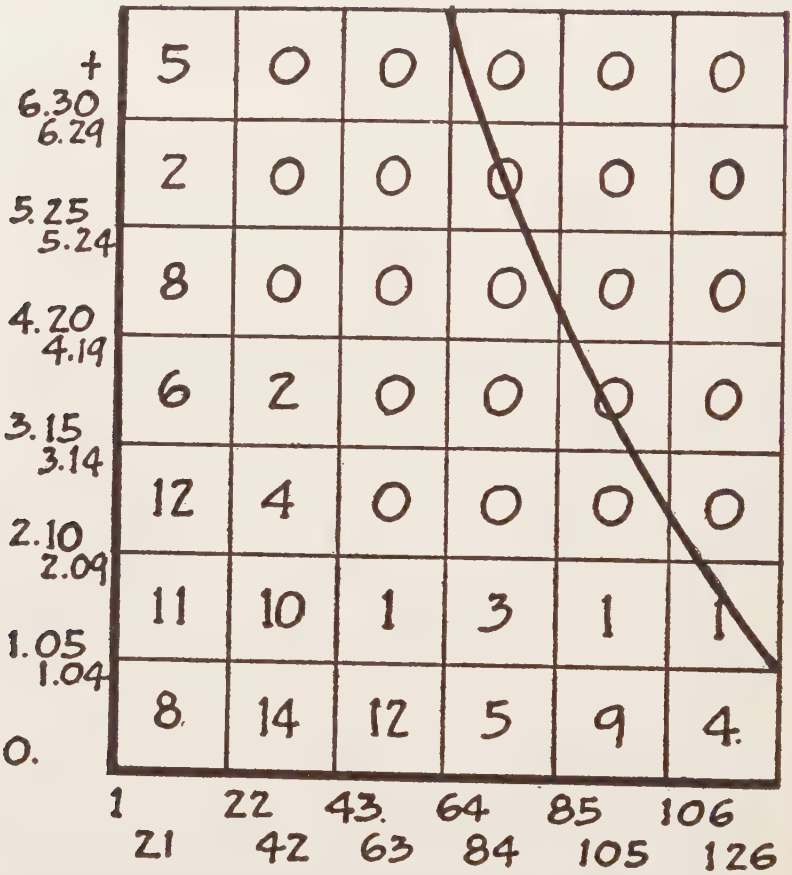

FIGURE 3. ASSOCIATION BETWEEN FREQUENCY AND DEVIATION VALUE OF RESPONSES

The horizontal axis indicates the frequency of response, the vertical, the deviation of the response value from the general mean. The line at the right indicates the maximum deviation value for the various response frequencies.

obtain. The values along this line were determined as follows: The C.S.C. Scores for the entire group were listed in order of size from highest to lowest. Then for a frequency of 20, the mean of the 20 highest was computed, and so on. The line rep-

resents the smoothed curve running through the means such as these.

The relationship between the deviation and the frequency in the preceding diagram is inverse and curvilinear. Where the frequency is high, the deviation value tends to reach its maximum, but this maximum is low. As the frequencies become smaller the deviation values tend to distribute themselves more widely and the discrepancy between the actual and the maximal values becomes larger. The diagram further illustrates how differentiation is lessened when the same response is made by a large proportion of the entire group. It may prove helpful also in the interpretation of the results of the attempt to locate the source of item coefficient unreliability, presented below.

EMPIRICAL STUDY OF THE RELIABILITY OF THE ITEM COEFFICIENT

Whatever are the defects mentioned above, they are hardly less apparent in one form or another in the usual coefficients of association or correlation. The item coefficient has, however, one relatively peculiar defect. It has, at least thus far, successfully thwarted attempts to determine through mathematical treatment, a measure of its reliability. An empirical treatment is therefore necessary. The results of such a treatment on a small scale, together with a theoretical consideration of the matter, are presented in the following pages.

The data employed consisted of:

1. The response validity values and the response frequencies for forty items of the American Council Completion Test as computed with 137 cases.
2. The item validity coefficients based on the results with these 137 cases.
3. 1 and 2 above, based on the first 68 cases of the entire group.
4. 1 and 2 above, based on the last 69 cases of the entire group.

These data are given in part in Tables IV and V.

First, the coefficient of correlation was computed for the forty pairs of coefficients for each item; that is, the coefficient based on the first 68 cases was correlated with that based on the last 69. The correlation coefficient proved to be only .279. The P.E. of the coefficient is .099.

Second, the question arose, "Does the reliability of the item

coefficient vary with the size of the coefficient?'' In order to answer this question the difference between the first group and the second group coefficients for each item was correlated with the size of the item coefficient based on the entire group. The coefficient was found to be .524. There is, then, an apparent tendency for low coefficients to be more reliable in absolute terms. This is to be expected from the fact that the smaller the item coefficient based on the entire group, the less is the possible range of difference between the item coefficients based on the halves of the group. When the difference between the first and second group coefficients was correlated with the size of the coefficient based on the first half group and not on the entire group, the r fell to .300. The conclusion still remains, although to a less striking degree, that with the data employed, good items are less reliable in absolute terms, than poor items. The P.E.'s of the coefficients are respectively .078 and .097.

The third approach involved an analysis of the component parts of the item coefficient formula and a consideration of the reliability of the various parts. The most significant factors of the formula are the response frequencies and the response values. Coefficients of correlation between the first and second groups for these factors were:

n values of the first half with the n values of the second half	$.975 \pm .005$
M values of the first half with the M values of the second half	$.341 \pm .095$

The most significant cause of unreliability seems to be contained in the response C.S.C. Score values. These values are largely dependent upon the reliability of the C.S.C. Scores themselves. Hence the results of this tentative study point especially to the need of determining empirically the relationship between the reliability of the item coefficient and the criterion score reliability.

In the fourth place, does the reliability of the response value vary with the size of the response group? Returning to the method of assigning C.S.C. Score response values (described on pages 14 ff) it seems probable that the response values of high frequency responses would show a higher reliability than those of low frequency. The results with the American Council Completion Test were studied to determine whether the above sup-

position was substantiated in fact. The difference between the response value with the first group and that with the second was employed as a measure of the reliability of the response value. The response frequency is simply the number of students out of the entire group making the given response. In Figure 4

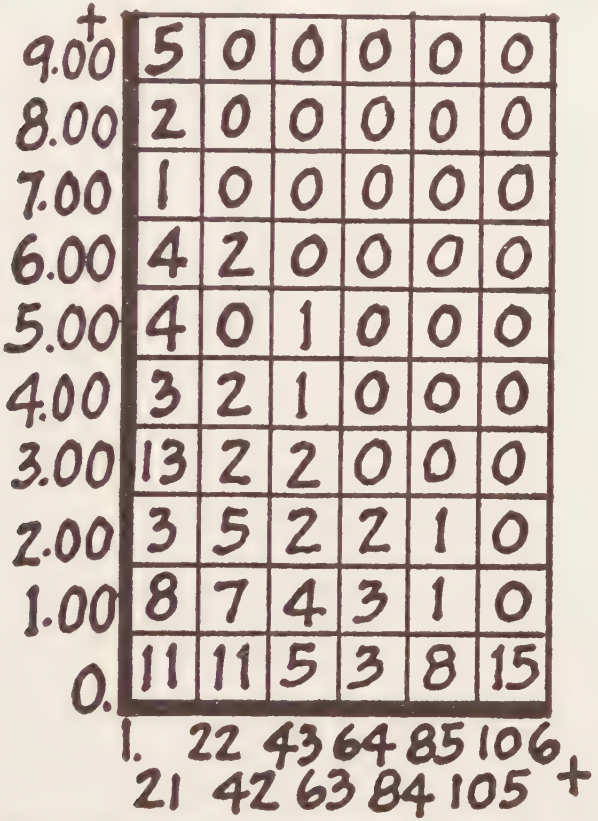

FIGURE 4. ASSOCIATION BETWEEN FREQUENCY AND RELIABILITY OF RESPONSE

The frequency of response is indicated on the horizontal axis, reliability of response along the vertical axis.

is plotted the frequency of occurrence of the differences between the first and the second group response values corresponding to the various sizes of the response group. As the response frequency increases, the mean of the differences and their variability decrease. With increasing frequencies, then, the

reliability of the response values becomes both greater and more constant. Consequently, it would appear reasonable to say that the item coefficient may be made more reliable by increasing the number of cases employed in its computation.

Fifth, how does the reliability of the response value vary with the size of the deviation of the response group mean from the mean of the entire group? The contingency diagram of Figure 5

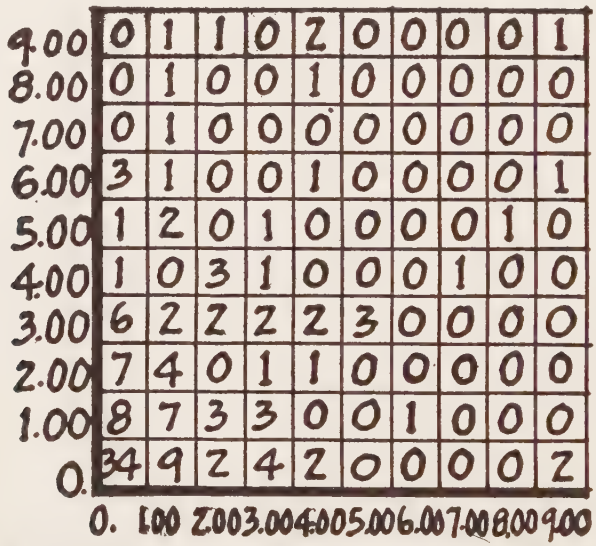

FIGURE 5. ASSOCIATION BETWEEN DEVIATION VALUE AND RELIABILITY OF RESPONSE

The horizontal axis indicates the deviation of the response value from the general mean; the vertical, the reliability of the response value.

presents the reliability as above plotted against the deviations of the response C.S.C. Score value from the mean C.S.C. Score of the entire group. Most of the cases are contained in the lower regions of both scales, and hence the association is high at the lower extremity. For the cases falling elsewhere, the association is very small. High deviations, then, in themselves do not indicate the probability of high reliability.

PRACTICAL EFFECTIVENESS OF THE ITEM COEFFICIENT IN CHOOSING THE BEST ITEMS

In the light of the high unreliability of the item coefficient as computed with the available data, it is evident that some proof

that the coefficient really selects the best items must be advanced before valid results may be found with the use of the coefficient concerning the factors associated with item goodness. The items of the American Council Completion Test were divided twice into four groups of ten, first on the basis of the item validity coefficient as computed with the first 68 cases, and second, on the basis of the coefficient as computed with the last 69 cases. The meaning of the "ten best items" and the "ten worst items" as used below is self-evident. The "first mediocre set" includes the items ranking sixteenth through thirty-fifth in size of item coefficient. The "second mediocre set" includes those ranking eleventh through fifteenth, and twenty-sixth through thirtieth. Scores according to both the author's scoring method and the new scoring technique were computed for each set of items. These scores were correlated with the C.S.C. Scores to give the validity of the set. The coefficients of correlation are presented in Table XIV.

TABLE XIV

VALIDITY COEFFICIENTS OF SETS OF ITEMS OF THE AMERICAN COUNCIL COMPLETION TEST ILLUSTRATING THE EFFECTIVENESS OF THE ITEM COEFFICIENT

Items	Selection Based on First 68 Cases		Selection Based on Last 69 Cases	
	Old Scoring	New Scoring	Old Scoring	New Scoring
Ten best items	.359 ± .071	.338 ± .073	.031 ± .081	.091 ± .080
Ten worst items	.139 ± .080	.042 ± .082	.050 ± .081	.135 ± .080
First mediocre set	.391 ± .069	.328 ± .073	.206 ± .078	.249 ± .076
Second mediocre set ..	.309 ± .074	.125 ± .081	.272 ± .075	.274 ± .075
All items	.340 ± .072	.345 ± .072	.161 ± .079	.232 ± .077

The coefficients of correlation are computed in each case with the group other than that used in the determination of the item validity coefficients and the new scoring values. In the case of the selection based on the first 68 cases, one of the mediocre sets proved somewhat higher than the best set in the size of the validity coefficient of correlation based on the old scoring method scores. In the case of the selection based on the last 69 cases, the ten best items yielded the lowest validity correlation for both the old and the new scoring method scores, the item coefficient being, then, insufficiently effective.

The old scoring and the new scoring coefficients retain about the same relative position for each set of items. This fact elimi-

nates the possibility of criticism of the use of the new scoring validity coefficients in evaluating the effectiveness of the coefficient.

The thirty items of the Roback Test 1 were divided into three sets of ten, according to the size of the item coefficient. In addition, the best five items were similarly selected. Scores were computed for these sets according to the new scoring method and were subsequently correlated with criterion scores, yielding the following validity coefficients:

Ten best items	$-.020 \pm .081$
Ten mediocre items	$-.147 \pm .079$
Ten worst items	$-.147 \pm .079$
Five best items	$.027 \pm .081$
All items	$-.148 \pm .079$

The scoring values and the item coefficients were based upon the first 68 cases. The coefficients of correlation were computed with the last 69 cases.

The item coefficient appears to have been successful in selecting the best five and, to a large extent, the best ten of the items, but has not adequately differentiated between the second and third item sets.

Similarly, a set of the best ten items and a set of mediocre items were selected out of the sixty items of the Thorndike Academic Information Test, the IIG8. The mediocre set included those ranking twenty-sixth through thirty-fifth in size of item coefficient. The item coefficient does not adequately distinguish between the best and the mediocre sets, the validity coefficients of correlation being as follows:

New Scoring Method Score with the Ten Best Items	$-.138 \pm .077$
New Scoring Method Score with the Ten Mediocre Items	$-.109 \pm .077$
New Scoring Method Score with All the Items	$-.011 \pm .078$

With the data employed the item coefficient has proved only moderately successful at best in differentiating between the effective and the ineffective items. This may be due to several causes, the most significant of which are listed below.

First, the items studied, having already been carefully selected by expert psychologists, might be expected to have relatively

restricted true differences in goodness, and hence further differentiation within the restricted range is made difficult.

Second, the factors making for the unreliability of the coefficient, such as the unreliability of the criterion scores and the smallness of the groups with which the coefficients were computed, tend to destroy the effectiveness of the item coefficient.

Third, the selection of items by means of the item *validity* coefficient necessarily disregards the intercorrelations among items. A tentative suggestion purporting to overcome this difficulty in part is made below.

Fourth, since the item coefficient was originally intended for a test (the McCall Multi-mental) in which the responses might be regarded as having degrees of value rather than as being quite entirely correct or incorrect as is the case with most intelligence test items, certain errors might have resulted. A modification of the item formula which allows for this situation is discussed below.

The first and second causes of ineffectiveness mentioned above may be partly eliminated by changes in the original selection of items for study and in the selection of subjects for study.

In connection with the third source of error, since it is entirely infeasible to compute the intercorrelations among items, a less involved substitute method is necessary. The theory underlying the present tentative suggestion is that once items have been grouped together as measuring the same trait, a rough approximation of the average intercorrelations for each item might be obtained by correlating each item with the total score on all items. Thus a "consistency" coefficient would be computed for each item. Paralleling the reasoning when true intercorrelations are employed in estimating the value of an item, it would become necessary to weight inversely the consistency, and to weight directly the validity of an item. Thus in selecting between two items of identical validity, the one having the lowest validity would be expected to be more effective when joined with other items to yield a test score. The value of this suggestion and more exact directions as to its use can be shown only after much intensive study. As an indication of this kind of trial and error research that seems necessary, the following is noted:

It was first necessary to compute the item consistency coefficients for the items of the American Council Completion Test,

employing the total score on the subtest as the criterion. Incidentally, these coefficients are more reliable than the validity coefficients since the reliability of the criterion scores is considerably higher than that of the C.S.C. Scores, as indicated by a reliability coefficient of .797, computed in the usual manner, i.e., by correlating half-test scores and estimating the whole-test score reliability coefficient by means of the Spearman-Brown formula. The ten best items of the American Council Completion Test were then selected on the basis of the item validity coefficient. The validity of this set of items when rescored by the new scoring method is represented by a coefficient of correlation of .519, computed with the cases on which the new scoring values were based. The coefficient was lowered to .460 when the items were selected so that they represented the worst ten in consistency of the best twenty in validity. This one result is necessarily inconclusive; the final solution of the matter is not attempted in the present study.

MODIFICATION OF THE ITEM COEFFICIENT

In connection with the fourth source of error indicated above, some evidence was found to lead to the conclusion that a modification of the item coefficient formula would prove beneficial. Often in the tests employed in the study it was found that the response which the author regards as incorrect would be found to yield a higher response value in terms of criterion scores than that which the author credits as correct. The coefficient as heretofore employed credits the differentiation between response values even where their direction is contrary to subjective logic. Whether logic should be upheld in the situation must be determined again through empirical results. The ten best items of the Thorndike IIG8 Test were selected, first, according to the usual coefficient, and, second, according to a coefficient described below, which decreases in size when the incorrect response value is higher than the correct response according to the author. The validity coefficients of correlation were:

Ten best items on basis of usual item coefficient...	$-.138 \pm .077$
Ten best items on basis of modified item coefficient	$.097 \pm .087$

In this instance, a marked improvement resulted from the use of the modified item coefficient.

The modified formula described on page 38 is written like the original:

$$C = \frac{(M_1 - M_2)\ (n_1 \times n_2) + (M_1 - M_3)\ (n_1 \times n_3) + (M_2 - M_3)\ (n_2 \times n_3) \ldots \text{etc.}}{\text{S.D. } N^2}$$

but the M_1, M_2, M_3, etc., are no longer necessarily in the order of size from highest to lowest. Where an adequate judgment, supported by whatever objective evidence is available, can be expected to indicate that a certain response is better than a second response, then the M value of the response said to be worse is subtracted from the M value of the response judged to be better, even though the worse has a higher value in criterion score terms than does the better. Certain "(M — M) (n × n)" terms may then become negative. The modified procedure operates on the assumption that the reversals of the response values from the logically expected order are errors which the item fails to avoid, and hence ought to lower the coefficient for the item. Where there is doubt as to the order of the response values, then the previous procedure is followed; namely, that which places first the M value which was actually found to be highest in terms of the criterion scores, and so on. The "(M — M) (n × n)" term for such cases is consequently always positive.

DETERMINATION OF THE OBJECTIVE FACTORS ASSOCIATED WITH ITEM GOODNESS

Because of the low reliability and effectiveness of the item coefficient as employed with the present data, the study of the relation of certain objective measures to the measure of item validity can retain but small significance. Such objective item analysis must await the proof of the effectiveness of the measure of item goodness. The treatment of this phase of the study is consequently brief and of a tentative nature.

The item validity and consistency coefficients for several of the subtests are presented in Table XV. Item numbers are listed vertically at the extreme left and in the case of the last twenty items of the Thorndike IIG8 Test, at the right.

Figure 6 represents graphically the data of Table XIV concerning the validity coefficients of the various items. The figure illustrates the significantly wide variability of the measured validity goodness of the items contained within a given subtest.

TABLE XV

ITEM VALIDITY COEFFICIENTS FOR SEVERAL SUBTESTS, WITH THE HALF GROUP ITEM VALIDITY AND THE WHOLE GROUP ITEM CONSISTENCY COEFFICIENTS FOR THE AMERICAN COUNCIL COMPLETION TEST

Item	American Council Completion Test				Brown 3 Validity	Roback 1 Validity	Thorndike IIG8 Validity		
	Validity		Consistency						
	Gr. 1.	Gr. 2.	All	All					
1	.076	.089	.080	.121	E .000	.071	.137		
2	.018	.017	.012	.084	.015	.123	.049		
3	.043	.051	.044	.134	.025	.077	.111		
4	.020	.057	.051	.165	.020	.107	.105		
5	.071	.095	.070	.131	.011	.058	.055		
6	.070	.114	.091	.158	.013	.042	.077		
7	.082	0.54	.044	.210	.039	.061	.108		
8	.051	.022	.035	.128	.089	.024	.109		
9	.091	.068	.079	.066	.007	.067	.079		
10	.045	.051	.034	.168	.080	.051	.055		
11	.034	.047	.037	.151	F .000	.077	.086	(Item)	
12	.040	.043	.024	.185	.000	.111	.036	41	.073
13	.063	.108	.086	.130	.012	.048	.048	42	.146
14	.037	.012	.011	.067	.010	.047	.059	43	.137
15	.073	.081	.051	.183	.021	.051	.057	44	.007
16	.011	.114	.056	.099	.001	.050	.066	45	.093
17	.041	.071	.078	.222	.027	.006	0.35	46	.181
18	.052	.007	.036	.193	.001	.070	.039	47	.165
19	.059	.054	.050	.140	.043	.058	.089	48	.066
20	.038	.014	.020	.199	.031	.123	.103	49	.087
21	.028	.017	.037	.253		.022	.062	50	.074
22	.060	.054	.056	.037		.083	.110	51	.068
23	.143	.073	.083	.216		.094	.032	52	.119
24	.084	.078	.065	.141		.093	.063	53	.119
25	.086	.091	.085	.088		.040	.122	54	.144
26	.157	.053	.090	.102		.083	.092	55	.170
27	.072	.111	.071	.165		.045	.044	56	.106
28	.173	.078	.124	.211		.092	.021	57	.062
29	.046	.093	.047	.087		.062	.171	58	.083
30	.086	.058	.048	.080		.115	.109	59	.114
31	.088	.162	.117	.211			.161	60	.098
32	.069	.116	.092	.172			.102		
33	.166	.121	.083	.219			.082		
34	.093	.085	.056	.258			.116		
35	.073	.164	.070	.246			.067		
36	.070	.128	.033	.158			.079		
37	.049	.051	.030	.264			.032		
38	.081	.114	.065	.265			.039		
39	.044	.160	.058	.230			.138		
40	.112	.073	.091	.116			.070		

It also indicates the difference in the average goodness and the variability of the goodness of the item coefficients taken as a group for each subtest. The standard deviations and the means of the item coefficients for each of the four subtests represented are as follows:

Test	S. D.	Mean	No. of Items	No. of Cases Employed
American Council 1 ...	26.595	.0625	40	137
Brown 3	23.695	.0218	20	100
Roback 1	29.115	.0674	30	68
Thorndike IIG8	38.875	.0873	60	100

The interrelationships between the validity, consistency, and difficulty, taken in pairs, were determined with the American Council Completion items. The item coefficients of the original form and as computed with all the 137 cases were used as

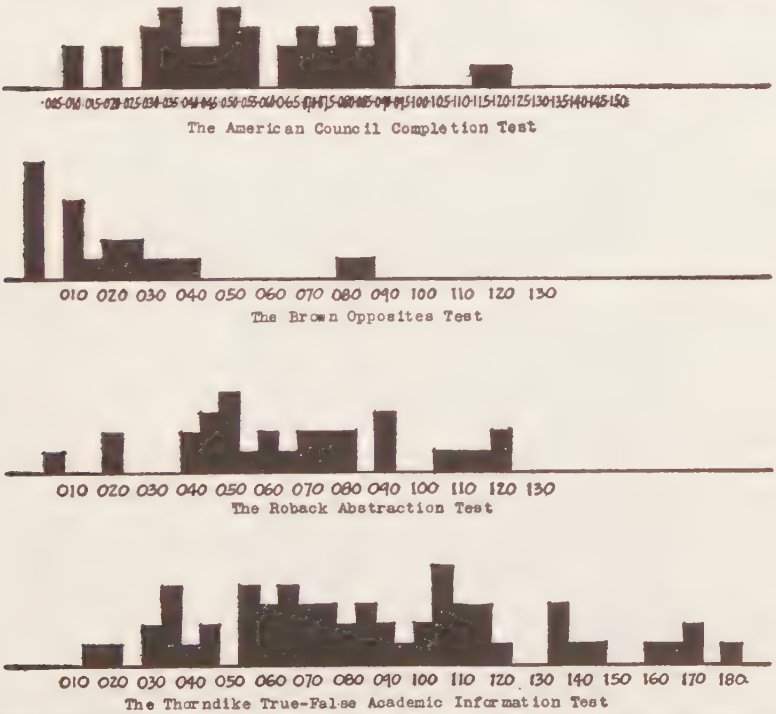

FIGURE 6. DISTRIBUTION OF THE ITEM VALIDITY COEFFICIENTS FOR SEVERAL SUBTESTS

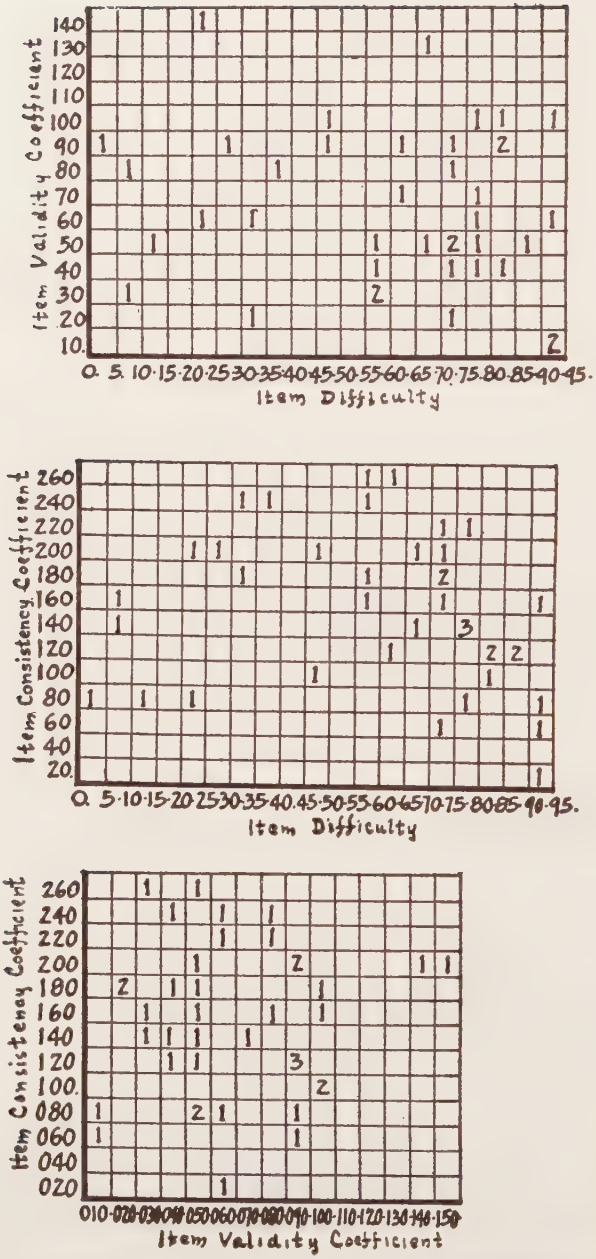

FIGURE 7. INTERRELATIONSHIPS BETWEEN VALIDITY, CONSISTENCY, AND DIFFICULTY OF ITEMS

measures of item validity and consistency. The detailed account of the determination of these coefficients is given on pages 38 and 49. The measure of difficulty employed was the percentage that the number making the correct response according to the author's scoring was of the total number making an attempt with the item. This method of calculating item difficulty eliminates somewhat the effect of the position of the item by omitting the group which apparently had insufficient time to reach the later items.

The contingency diagrams of Figure 7 indicate the interrelationships. The validity measures are practically uncorrelated with the consistency or the difficulty measures. The consistency and the difficulty measures appear to have a slight curvilinear relationship; the middle ranges of difficulty are associated in general with low consistency. This last phenomenon may be due to the fact that the item coefficient tends to penalize extreme difficulties, as indicated on page 40. That the same relationship was not found in the case of validity with difficulty may be due to the greater unreliability of the validity coefficients (pages 10 and 50).

CHAPTER V

THE ANALYSIS OF THE SUBTESTS

The treatment of subtests as if they were true elements of a psychological examination is defective, theoretically, essentially because the component items are dissimilar in their power of differentiation and also, to a less certain extent, in their true effective or "psychological" content. However, the analysis of subtests treated as units has several values.

First, it is an essential step in item analysis in that it aids in the adequate selection of items for intensive analysis and in that it gives the predictive value of the subtest, utilizing the author's scoring key, against which may be compared the predictive values of the subtest according to other methods of scoring. This use is illustrated and developed in the discussion in Chapters III and IV.

Second, it is valuable in itself as indicating roughly the degree to which factors such as reliability, difficulty, form, and the like, are associated with the validity of various types of test items. This second use is limited, and for two reasons. First, the items within any subtest vary so much as to make the crude summary of those items which is represented by the subtest score simply suggestive of the probable true relationships, and nothing more. Second, the subtests, to a far greater extent than the items, are unequated for irrelevant factors. The results of the study of the subtests, which is the subject of this chapter, are partially invalidated by the limitations indicated above.

DETERMINATION OF THE MEASURES EMPLOYED

The measure of validity of a subtest was determined by computing the Pearson coefficient of correlation between the scores on that test and the C.S.C. Scores. The method of determining the criterion scores is explained in Chapter II, pages 7 ff. In the case of the Brown and Roback subtests, the raw scores were transmuted into T scores. This transmutation does not effect the

relative positions of the scores and hence modifies the computed coefficients of correlation only to a negligible extent. All but five of the thirty-eight subtests which were administered yielded scores sufficiently variable for the computation of the validity measure.

The reliability measure was determined by first dividing the subtest into two equal or practically equal groups of items, second, summing the scores on each half-test, and, third, correlating the score on one half of the test with that on the other. The usual assumption was made that the values of the test are similarly matched. The resultant coefficient gives, then, the reliability or consistency of the half-test. Twelve of the subtests did not yield coefficients, because they were either inadequate in differentiation or serial in nature.

The testing time is used as an approximate measure of the number of minutes spent by the student in responding to the item stimulations. Slight inaccuracies in this measure, such as that caused by the fact that the faster student may have spent only part of the time in actual work or that due to the flexible time limits characteristic of the Brown tests, were of necessity neglected.

Difficulty was determined by dividing the mean score on a test by the possible maximum score on that test.

The speed of item response gives for each test the time spent per item, taken on the average.

The form of the test items refers simply to the number of responses that may be attempted. Thus Type A includes two or three choice test items; Type B, four through eight choice items; Type C, practically unlimited choice items; and Type D, unclassifiable items.

The so-called content types are as follows:

1. Language: comprehension.
2. Language: manipulation.
3. Language: comprehension and use—completion.
4. Language: analogies.
5. Language: opposites.
6. Judgment and reasoning.
7. Information.
8. Numerical or algebraic manipulation.
9. Special material.

TABLE XVI

DATA EMPLOYED IN THE ANALYSIS OF THE SUBTESTS

Subtest	Val. Coeff. and P. E.		Rel. Coef. and P. E.		Time	Difficulty	Time per Item	Form	Content
BROWN	(N 100)		(N 248)						
1. Completion	.230	.063	.287	.039	13	39	.65	C	3
2. Vocabulary	.250	.063	.755	.018	7	68	...	D	1
3. Opposites	.162	.065	.211	.041	8	84	.40	B	5
4. Analogies	.179	.065	.361	.037	8	83	.40	B	4
5*c*. Arith. Comput. ..	.101	.066	.356	.038	12	79	.60	C	8
5*r*. Arith. Reasoning.	.169	.065	.463	.034	12	44	.60	C	6
AMERICAN COUNCIL	(N 137)		(N 137)						
1. Completion	.252	.054	.662	.033	10	55	.25	C	3
ROBACK	(N 137)		(N 137)						
1. Abstraction	.020	.058	.482	.045	10	52	.33	C	2
2. Analogy	.089	.058	.512	.043	10	63	.40	C	4
3. Relations	.100	.057	.202	.056	15	70	1.00	C	6
4. Insertion	.185	.056	....		15	58	...	D	2
5. Reference	.109	.057	.347	.051	15	29	1.00	C	2
6. Opposites	.052	.058	.388	.049	10	40	.40	C	5
7. Subsumption	.114	.057	.336	.051	20	41	1.67	C	2
8. Directions	.111	.057	....		5	82	...	D	1
9. Judgment	.040	.058	.396	.049	15	55	.50	B	6
10. Cryptogram	.147	.057	....		15	78	...	D	6
THORNDIKE	(N 175)		(N 175)						
Part I									
1. Easy Directions..	....		....		3	95	.33	D	1
2. Mixed Sentences.	....		.163	.050	5	87	.25	A	2
3. Arith. Comput. ..	.099	.050	.121	.050	4	67	.25	C	8
4. Arith. Reason. ...	.234	.048	.483	.039	8	65	.40	C	6
5. Information	.089	.051	.229	.048	4	72	.20	B	7
6. Opposites	.219	.049	.484	.039	4	82	.10	A	5
7. Best Answer	....		....		4	94	.40	B	6
8. Number Completion	.168	.050	.672	.028	4	77	.17	C	8
9. Analogies	.137	.050	.539	.036	4	80	.10	A	5
10. Large and Small Numbers	....		....		4	92	.20	D	8
11. False Statement..	....		....		4	60	.25	A	6
12. Reasoning	.108	.050	.519	.037	4	50	.25	A	6
13. Symbol Learning.	.178	.049	.404	.043	8	90	.20	D	9

TABLE XVI (*Continued*)

Subtest	Val. Coeff. and P. E.		Rel. Coef. and P. E.		Time	Difficulty	Time per Item	Form	Content
Part II									
3. Missing Parts ...	.056	.051	....		3	62	.30	D	9
4. Picture Analogies	.189	.049	....		3	65	.37	B	9
5. Geom. Figure Anal.	.143	.050	....		3	86	.30	C	9
6. Algebra	.213	.049	....		10	75	1.67	C	8
7. Mech. Information	.027	.051	....		4	13	.40	C	7
8. Gen. Acad. Infor.	.099	.050	.132	.050	13	28	.22	A	7
Parts II and III									
1. Reading	.168	.050	.371	.044	36	43	...	C	1
2. Language Completion	.148	.050	.536	.036	28	26	1.17	C	3

RESULTS

Table XVI presents for the subtests studied all the available measures of validity, reliability, difficulty, time per item, form, and content. Time is given in terms of minutes.

The lowness of the validity coefficients as compared with those usually reported may be due in part to the unreliability of the criterion scores, in part to the fact that an unusually large number of students engage in outside work, and in part to the restricted variability of the group.

In studying the relationship between validity, reliability, and time, taken in pairs, rank method coefficients of correlation were computed, utilizing the twenty-five complete sets of measures of these functions. The *r* values, (transmuted from rho) are:

Validity with Reliability ..... .403 ± .113
Validity with Testing Time ..... —.147 ± .132
Reliability with Testing Time ..... —.284 ± .134

Although factors other than time are probably the causes of the inverse variation, the indications are that the time devoted to a test is not a very significant factor in causing high validity or high reliability. In general, the more reliable test is the more valid test, but it is difficult to say whether that is due to the transference of reliability to validity or to the fact that a given

test author has selected tests according to a standard that is equally high, relatively, for both.

When the effect of testing time is held constant, the coefficient of correlation of validity with reliability is found, by the use of partial correlations, to be .381.

Difficulty and time per item were each correlated with validity, yielding the following product-moment coefficients:

Validity with Difficulty	$-.327 \pm .104$	N 33
Validity with Time per Item	$.001 \pm .127$	N 28

In general, the less difficult the test, the more valid it proved.

The mean validity coefficient for the subtests grouped according to type of test form are as follows:

Type	Mean Validity Coefficient	No. of Tests
A	.143	3
B	.133	6
C	.135	18
D	.155	6

The multiplicity of choice as an index of test form shows no significant association with the validity of the subtests.

The association between test content as analyzed and test validity is more striking, as indicated by the following means of the subtest validity coefficients for the various content types:

Type	Mean Validity Coefficient	Rank	No. of Tests
1	.176	2	3
2	.107	8	4
3	.210	1	3
4	.135	6	3
5	.145	3.5	3
6	.130	7	6
7	.072	9	3
8	.145	3.5	4
9	.141	5	4

These results are conditioned by errors of classification, by the unreliability of the item coefficients, by the operations of "irrelevant" factors, such as test author, length of test, and so on, and by chance influences. It is interesting to note, however, that certain usual findings such as the superiority of the language completion type and the inferiority of the information type, are substantiated. (See page 57 for key to types.)

CHAPTER VI

SUMMARY AND CONCLUSIONS

1. The purpose of the present study is to ascertain whether college entrance intelligence tests may be improved by the use, in connection with the selection and scoring of test items, of certain relatively objective statistical devices.

2. The relatively objective new scoring method employed bases the value of a response to an item stimulation upon a measure representative of the College Success Criterion Scores earned by those students who have made a given response. The conclusions numbered 3 through 9 are based on the use of the mean as this representative measure.

3. Test constructors usually assign a single scoring value to a certain type of response such as the subjectively determined "correct" response, whereas, when evaluated in terms of the College Success Criterion Scores, any given response type shows, for the various items, a wide distribution of values.

4. Within any one item, various types of response, such as an incorrect attempt and an omission, are very often assigned the identical scoring value, whereas the more objective measure here employed usually indicates different values for the different types of response.

5. In general, to omit a response to an item stimulation is an indication of higher College Success Criterion Score than it is to make an incorrect attempt.

6. In the majority of the tests the few students who were outstandingly slow in their test reactions proved, in general, to be markedly superior in College Success Criterion Score.

7. The possibility of correcting the test scoring key by means of objective item response analysis is illustrated in the case of responses which the author's scoring key regarded as wrong, but which the scorer thought deserved credit, and some of which were made by students with high College Success Criterion Scores, on the average.

8. The empirical comparison of the old and the new scoring method indicates results which with two tests point slightly favorably to the new technique, with two others, slightly against it, and with two others markedly against it.

9. The indications are that the new scoring method failed to produce any significant improvement, essentially because of the lowness of the original correlation between the tests and the criterion scores on which the new scoring values were based and because of the dissimilarity of the group employed in determining the new scoring values with the group with which the values were tried in the rescoring of the tests. The new technique must for practical purposes prove its worth with groups as dissimilar as those here employed, but it ought to receive a trial with tests showing higher validity coefficients than those used here. This implies experimentation in a field where criterion scores that are highly reliable can be found.

10. A tentative investigation to determine whether some value other than the mean ought not to be employed as representing the College Success Criterion Scores associated with a given item response, revealed for a single set of items, the following order of merit of the various measures, from best to poorest: the median, the upper quartile, the mean, and the proportion of a superior group making a given response. Each of these measures, for the same set of items, proved inferior to the old scoring method of the author.

11. There is evidence to show that an approach toward the uniform weighting of the items of a set, achieved by limiting the distance between the lowest and the highest item response value to a given number of units, is an improvement over the weighting of items in proportion to their validity differentiating power.

12. The analysis of the characteristics of the item coefficient invented by McCall and modified by others, indicates that the measure is theoretically sound where chance errors are minimized. Its one significant defect is the fact that an associated reliability measure cannot, apparently, be devised algebraically.

13. The reliability of the coefficient as computed empirically by correlating item coefficients determined with the same items but with two different groups is represented by a coefficient of correlation as low as .279.

14. The higher item coefficients prove no more reliable than the lower ones, when the reliability of the coefficient is measured in terms of the absolute difference between the first group and the second group item coefficients.

15. There is evidence to show that the unreliability of the coefficient is due more to the unreliability of the response values than to the unreliability of the response frequencies. Since these values are largely dependent upon the College Success Criterion Scores, the original source of the item coefficient unreliability may be traced in the last analysis to the unreliability of the criterion scores.

16. While high response frequencies are in general associated with high reliability of response value, the size of the response value is unassociated with response value reliability.

17. The item coefficient is only moderately successful at best in selecting sets of items that are the ten best, the ten poorest, and so on.

18. A single tentative trial of the use, in selecting the best items, of a "consistency" item coefficient along with the validity item coefficient, failed to improve the selection.

19. The writer's modification of the item coefficient, as described, and as employed with a single test, resulted in a marked improvement in the selection of the ten best items.

20. The item validity goodness, as measured by the item coefficient, shows a significantly wide variability for the items contained in any given subtest.

21. The variability and the mean of the measures of item goodness vary for the several subtests studied.

22. The measure of item validity goodness is practically uncorrelated with either that of item consistency or that of item difficulty.

23. The consistency and difficulty measures show a slight curvilinear relationship, the middle ranges of difficulty being associated with the highest consistency, while the extremities in difficulty are associated, in general, with low consistency.

24. The coefficient of correlation (rho transmuted into r) between the validity and the reliability coefficients for twenty-five subtests proved to be .403. When the time of the tests is held constant (by means of partial r's) the coefficient of correlation of validity with reliability is slightly lowered.

25. Testing time shows a low negative correlation with validity, and with reliability, a somewhat higher negative correlation. These negative correlations may very likely be due to causes associated with the variables, such as the fact that subtests by the same author tend to have similar testing times.

26. The validity of subtests shows a fair inverse correlation with their difficulty.

27. The validity of subtests and the average time per item for the subtests are uncorrelated.

28. The validity of subtests and the multiplicity of choice as an index of test item form, show no significant association.

29. The evidence indicates a fairly high association between the content of subtests and their validity. The language completion type, for example, excels the other types, while the information type is lowest of the nine categories.

The above conclusions are conditioned by various unreliabilities, as indicated in the body of the report.

Although certain relatively minor conclusions seem to offer definite justifications for the employment of objective item analysis, the results of the present study fail to realize the hope for the outstanding improvements that it was thought would grow out of such an analysis. The defeat of highly objective and intensive item analysis may be largely attributed to the inadequacy of the criterion scores employed.

In so far, then, as tests show low correlations with criteria; in so far as criterion scores show low reliability coefficients; in so far as it is necessary to employ a small number of cases in determining new scoring values; and in so far as basal groups (that is, groups with which the new scoring key is determined) differ from trial groups (or groups with which the determined scoring keys are to be used)—so far is the probable value of the empirical method limited.

It is reasonable to presume that with increased validity coefficients, reliability coefficients, numbers of experimental cases, and similarity of groups, the objective method will yield sufficient improvement to warrant the additional expenditure of effort entailed. The determination of "critical points" of the factors enumerated, below which the new technique fails to result in improvement, opens a field for needed and important investigation.

APPENDIX I

SUGGESTIONS FOR DECREASING THE LABOR ASSOCIATED WITH ITEM ANALYSIS

The use of certain forms and procedures may decrease considerably the time and labor connected with the analysis of items. The following are suggestions which have grown out of the writer's experience with this type of work.

1. In the selection of symbols for the various types of responses, the test author's credited responses should be indicated by corresponding numerals, where possible; responses assigned a value of zero should be represented by other symbols. This will facilitate the determination of the author's scores.

2. Where the scoring is not too complicated, the tabulation of the response symbols should take place simultaneously with the scoring of the tests. With the use of the tabulation form illustrated on page 12, this will entail no great hardship.

3. The computations involved in the determination of the new response values, when the mean of the criterion scores of the response group is used, may be made most economically as follows:

(1) Transmute the criterion scores of the entire group into plus and minus deviations from an assumed mean of the scores.

(2) Compute the sum of the plus and the sum of the minus deviations of these scores.

(3) Add the sums algebraically.

(4) On a narrow strip of paper, place next to the number representing each student, his criterion score deviation; in one color, if plus or zero; in another, if minus. The student numbers and criterion score deviations should be placed on the slip so as to correspond with the tabulation of the response symbols illustrated on page 12.

(5) To tabulate the criterion scores associated with each response, the criterion score deviation slip should then be placed immediately adjoining the tabulation of the response symbols for the item to be studied.

(6) The most frequent response type should be determined through inspection and omitted from the tabulation of associated criterion scores.

(7) Add algebraically the criterion score deviations for each of the responses tabulated.

(8) Divide the results of (7) by the respective response frequencies to obtain the response value indicated as the deviation in terms of criterion score units, of each of the response group means from the *assumed* mean

of the entire group. The value of the most frequent response will not as yet have been determined.

(9) Join into one algebraic sum the results of (7), computed for each of the tabulated responses.

(10) Change the sign of the total sum of (9).

(11) Add algebraically the result of (3) with that of (10).

(12) Divide the result of (11) by the frequency of the omitted response group to give the deviation of that response group mean from the *assumed* mean of the entire group.

(13) Divide the result of (3) by the frequency of the entire group. This will give the unit correction indicating the difference between the true and the assumed means.

(14) Subtract from each response group result indicated in (8) and (12) the unit correction of (13). This will yield the values of the various responses expressed as deviations in terms of college success criterion score units, of the response group means from the true mean of the entire group.

4. In rescoring responses and in computing the item coefficient, since distances between the response values is desired, it is not necessary to compute steps (13) and (14) of the above analysis.

5. In rescoring the responses, it was found best to repeat the tabulations illustrated on page 12, merely substituting the new values for the old.

6. In rescoring responses, it is helpful to consider the most frequent response as zero, and to determine the others according to the differences of each from the most frequent response value.

7. The modification of the scoring method is helpful in two cases: first, where a very extreme value is determined with a very small group, and second, where a student fails to attempt a good number of the later items, presumably because of lack of time. To avoid the errors and difficulties inherent in these two situations, extreme values earned by three or less students were reduced according to a set scale, and where a student failed to attempt a number of items at the end of the test, the lowest value of the responses of each item was assigned in the case of each item omitted.

APPENDIX II

MISCELLANEOUS SUPPLEMENTARY RESULTS

1. Total scores for each of the examinations were obtained by summing the scores on the several subtests contained in each, no attempt being made to weight the various subtest scores. The total examination scores were correlated with the College Success Criterion Scores to yield the following Pearson coefficients:

Examination	Val. Coef.	P.E.	No. of Cases	Group
Brown University	.265	.062	100	B
Roback	.111	.057	137	C
Thorndike	.235	.048	175	A
Thurstone IV	.276	.079	61	D

The D Group was more variable than the other groups, and hence a small downward correction of its validity coefficient is necessary. The lowness of the coefficients as compared with those found elsewhere is due in part to the factors indicated on page 59.

2. The validity coefficient of the Thorndike Examination rises from .235 to .290 when the score of the IIG8 test, that is, the true-false academic information test, is omitted from the composite.

3. The Reading Test items of the Thorndike Examination fall readily into four divisions. The reliability of the quarter-tests was determined by computing the Pearson coefficient of correlation between each possible pair of quarter test scores, and obtaining the mean coefficient associated with each quarter test. The results follow:

Correlation between Quarter-tests

	II 1 *a*	II 1 *b*	III 1 *a*	III 1 *b*
II 1 *a*	...	.303	.317	.291
II 1 *b*	.303	...	.178	.179
III 1 *a*	.317	.178	...	.076
III 1 *b*	.271	.179	.076	...
Mean	.297	.220	.190	.175

Portions of material apparently highly similar show significant differences in reliability. The results illustrate the need for the careful use and interpretation of the reliability coefficient.

4. The fluctuation of the results of the comparison of various scoring methods when few cases are employed, was brought to light by the compu-

tation of validity coefficients first with 35 cases, then with 40, and finally with the combined 75. The results follow:

Scoring and Selection	First 35	Second 40	Combined 75
Old Scoring—all items	.332	.121	.206
New Scoring—all items	—.336	.148	—.011
New Scoring—best 10 in validity coefficient	—.168	.014	—.138
New Scoring—best 10 in modified val. coef.	.035	.295	.097

The test employed was the Thorndike IIG8. The new scoring values were determined on a separate group of 100 students. It is apparent that the measured value of the several instances of scoring and selection methods varies considerably for the two smaller groups.

APPENDIX III

BIBLIOGRAPHY

ABELSON, HAROLD H. ('25) "Psychological Tests Versus High School Marks in Power of Predicting College Success." (Unpublished thesis, Teachers College, Columbia University.)

ANDERSON, J. E. AND SPENCER, L. T. "The Predictive Value of the Yale Classification Tests." *School and Society*, Vol. 24, p. 305.

BAILOR, E. M. ('24) *Content and Form in Tests of Intelligence.* Teachers College, Columbia University.

BROWN, WM. M. ('24) "A Study of the Predictive Value of Certain Kinds of Scores in Intelligence Tests." *Jour. of Educ. Psych.*, Vol. 15, p. 448.

CHAPMAN, J. CROSBY AND DALE, A. BARBARA. ('22) "A Further Criterion for the Selection of Mental Test Elements." *Jour. of Educ. Psych.*, Vol. 13, p. 267.

FOSTER, R. R. AND RUCH, G. M. ('27) "On Correction for Chance in Multiple-Response Tests." *Jour. of Educ. Psych.*, Vol. 18, p. 48.

GARRETT, HENRY E. *Statistics in Psychology and Education.* Longmans, Green and Company.

GATES, A. I. ('23) "The Correlation of Achievement in School Subjects with Intelligence Tests and Other Variables." *Jour. of Educ. Psych.*, Vol. 13, p. 223.

GATES, A. I. AND LASALLE, J. ('24) "Relative Predictive Values of Certain Intelligence and Educational Tests Together with a Study of Educational Achievement upon Intelligence Test Scores." *Jour. of Educ. Psych.*, Vol. 15, p. 517.

GEYER, DENTON L. ('23) "A Uniform Objective Examination of Intelligence Testing." *Jour. of Educ. Psych.*, Vol. 14, p. 372.

HERRING, JOHN P. ('25) "The Nature of Intelligence." *Jour. of Educ. Psych.*, Vol. 16, p. 505.

HOLZINGER, KARL J. ('24) "On Scoring Multiple-Response Tests." *Jour. of Educ. Psych.*, Vol. 15, p. 445.

JORDON, A. M. ('23) "The Validation of Intelligence Tests." *Jour. of Educ. Psych.*, Vol. 14, pp. 348, 414.

KELLEY, TRUMAN L. ('23) *Statistical Method.* The Macmillan Company.

LAIRD, DONALD A. ('24) "A Note on the Shortening of Examinations." *Jour. of Educ. Psych.*, Vol. 15, p. 116.

MAY, MARK A. ('23) "Predicting Academic Success." *Jour. of Educ. Psych.*, Vol. 14, p. 429.

MCCALL, WILLIAM A. *How to Experiment in Education.* The Macmillan Company.

McCall, William A. and his Students. ('26) "Construction of the Multi-mental Scale." *Teachers College Record,* Vol. 27, p. 394.

McCall, William A. and his Students. ('25) "The Multi-mental Scale." *Teachers College Record,* Vol. 17, p. 109.

Miller, George F. ('25) "Formulas of Scoring Tests in Which the Maximum Amount of Chance is Determined." *Jour. of Educ. Psych.,* Vol. 16, p. 304.

Ogden, Robert M. ('25) "The Nature of Intelligence." *Jour. of Educ. Psych.,* Vol. 16, p. 361.

Orleans, Jacob S. ('26) *A Study of the Nature of Difficulty.* Teachers College, Columbia University.

Otis, A. S. *Statistical Method in Educational Measurement.* World Book Co., Yonkers, N. Y.

Otis, A. S. *Directions: Otis Correlation Chart.* World Book Company, 1922.

Pearson, Karl. *Tables for Statisticians and Biometricians,* Part I, University College, London. 1924.

Pintner, R. ('26) "An Empirical View of Intelligence." *Jour. of Educ. Psych.,* Vol. 17, p. 608.

Pintner, R. ('26) "Accuracy in Scoring Group Intelligence Tests." *Jour. of Educ. Psych.,* Vol. 17, p. 470.

Pintner, R. ('23) *Intelligence Testing.* Henry Holt and Company.

Rice, A. R. ('25) "The Distribution of Intelligence among College Students." *Jour. of Educ. Psych.,* Vol. 16, p. 124.

Ruch and Koeth. ('23) "Power vs. Speed in Army Alpha." *Jour. of Educ. Psych.,* Vol. 14, p. 193.

Ruch, G. M. and De Graff, M. H. "Corrections for Chance and 'Guess' vs. 'Do not Guess' Instructions in Multiple-Response Tests." *Jour. of Educ. Psych.,* Vol. 17, p. 368.

Ruch, G. M. and Stoddard, G. D. "Comparative Reliability of Five Types of Objective Examination." *Jour. of Educ. Psych.,* Vol. 16, p. 89.

Rugg, Harold O. ('17) *Statistical Methods Applied to Education.* Houghton Mifflin Company.

Symonds, Percival. ('26) "Variations of the Product-Moment (Pearson) Coefficient of Correlation." *Jour. of Educ. Psych.,* Vol. 17, p. 458.

Symposium ('21) "Intelligence and Its Measurement." *Jour. of Educ. Psych.,* Vol. 12, Nos. 3 and 4.

Thorndike, E. L. ('14) *Educational Psychology,* Vol. III. Teachers College, Columbia University.

Thorndike, E. L. ('25) "The Improvements of Mental Measurements." *Jour. of Educ. Research,* Vol. II, No. 1.

Thurstone, L. L. ('25) "The Psychological Test Program." *The Educ. Record,* Vol. 7, No. 2.

Toops, Herbert A. ('26) "The Status of University Intelligence Tests in 1923-24." *Jour. of Educ. Psych.,* Vol. 17, p. 23.

Vincent, Leona. ('24) *A Study of Intelligence Test Elements.* Teachers College, Columbia University.

WEIDEMANN, CHAS. W. ('26) *How to Construct the True-False Examination.* Teachers College, Columbia University.

WILSON, WM. R. ('24) "Information as a Measure of Intelligence and Maturity." *Jour. of Educ. Psych.*, Vol. 15, p. 309.

WOOD, BEN D. *Measurement in Higher Education.* World Book Company, Yonkers, N. Y.

WOOD, E. P. ('27) "Improving the Validity of Collegiate Achievement Tests." *Jour. of Educ. Psych.*, Vol. 18, No. 1.

YULE, G. U. ('22) *An Introduction to the Theory of Statistics.* Chas. Griffin and Company, Ltd., London, England.